SKRIPSI

ANALISIS *CLUSTER* DENGAN METODE *ENSEMBLE ROCK*
UNTUK DATA BERSKALA CAMPURAN KATEGORIK DAN NUMERIK
(Kasus: Mahasiswa Aktif Program Studi Statistika FMIPA UNM)

NUR ARISKA

PROGRAM STUDI STATISTIKA
FAKULTAS MATEMATIKA DAN ILMU PENGETAHUAN ALAM
UNIVERSITAS NEGERI MAKASSAR
2017

SKRIPSI

ANALISIS *CLUSTER* DENGAN METODE *ENSEMBLE ROCK*
UNTUK DATA BERSKALA CAMPURAN KATEGORIK DAN NUMERIK
(Kasus: Mahasiswa Aktif Program Studi Statistika FMIPA UNM)

Diajukan kepada Program Studi Statistika Fakultas Matematika dan Ilmu Pengetahuan Alam Universitas Negeri Makassar untuk memenuhi salah satu syarat memperoleh gelar Sarjana Statistika

NUR ARISKA
1317142010

PROGRAM STUDI STATISTIKA
FAKULTAS MATEMATIKA DAN ILMU PENGETAHUAN ALAM
UNIVERSITAS NEGERI MAKASSAR
2017

HALAMAN PENGESAHAN

Skripsi ini diajukan oleh **Nur Ariska** dengan Nomor Induk **1317142010,** berjudul **Analisis *Cluster* dengan Metode *Ensemble ROCK* untuk Data Berskala Campuran Kategorik dan Numerik (Kasus: Mahasiswa Aktif Program Studi Statistika FMIPA UNM).** telah dipertahankan dihadapan dewan penguji dengan SK No. 4678/UN36.1/KM/2017, tanggal 18 Desember 2017 untuk memenuhi sebagai bagian persyaratan memperoleh gelar Sarjana Statistika pada Program Studi Statistika Fakultas Matematika dan Ilmu Pengetahuan Alam, Universitas Negeri Makassar pada hari Rabu Tanggal 20 Desember 2017.

Disahkan oleh:
Dekan Fakultas MIPA
Universitas Negeri Makassar

Prof. Dr. Abdul Rahman, M.Pd.
NIP. 196204171988031001

Panitia Ujian

1.	Ketua Ujian	:	Prof. Dr. Abdul Rahman, M.Pd.	(.....)
2.	Sekertaris Ujian	:	Muhammad Kasim Aidid, S.Si., M.Si.	(.....)
3.	Pembimbing I	:	Drs. Muhammad Nusrang, M.Si.	(.....)
4.	Pembimbing II	:	Sudarmin, S.Si., M.Si.	(.....)
5.	Penguji I	:	Prof. Drs. H. M. Arif Tiro, M.Pd., M.Sc., Ph.D.	(.....)
6.	Penguji II	:	Adiatma, S.Pd., M.Si.	(.....)

PERNYATAAN KEASLIAN

Saya bertanda tangan di bawah ini menyatakan bahwa skripsi ini adalah hasil karya sendiri, dan semua sumber yang dikutip ataupun yang dirujuk telah saya nyatakan dengan benar. Bila dikemudian hari ternyata pernyataan saya terbukti tidak benar, maka saya bersedia menerima sanksi yang ditetapkan oleh FMIPA UNM MAKASSAR.

Yang membuat pernyataan:

Nama : Nur Ariska
NIM : 1317142010
Tanggal : 20 Desember 2017

PERSETUJUAN PUBLIKASI UNTUK KEPENTINGAN AKADEMIK

Sebagai sivitas akademika Universitas Negeri Makassar, saya yang bertanda tangan dibawah ini

Nama : Nur Ariska
Nim : 1317142010
Program Studi : Statistika
Fakultas : Matematika dan Ilmu Pengetahuan Alam

demi pengembangan ilmu pengetahuan, saya menyetujui untuk memberikan kepada Universitas Negeri Makassar **Hak Bebas Royalti Noneksklusif (*Non-exclusive Royalty-Free Right*)** atas skripsi saya yang berjudul:

Analisis *Cluster* dengam Metode *Ensemble ROCK* untuk Data Berskala Campuran Kategorik dan Numerik (Kasus: Mahasiswa Aktif Program Studi Statistika FMIPA UNM).

beserta perangkat yang ada (jika diperlukan). Dengan Hak Bebas Royalti Non eksklusif ini Universitas Negeri Makassar berhak menyimpan, mengalih-media/format-kan, mengelola dalam bentuk pangkalan data (*database*), merawat, dan mempublikasikan skripsi saya selama tetap mencantumkan nama saya sebagai penulis/pencipta dan sebagai pemilik Hak Cipta, serta tidak dikomersialkan.

Demikian pernyataan ini saya buat dengan sebenarnya.

Dibuat di : Makassar
Pada Tanggal : 29 Desember 2017

Yang Menyatakan,

Nur Ariska

Menyetujui,

Pembimbing I	Pembimbing II
Drs. Muhammad Nusrang, M.Si.	**Sudarmin, S.Si., M.Si.**
NIP. 19661231 199103 1 020	NIP. 19701018 199703 1 001

MOTTO DAN PERSEMBAHAN

Mereka menjawab, " Mahasuci Engkau, tidak ada yang kami ketahui selain apa yang telah Engkau ajarkan kepada kami. Sungguh, Engkaulah Yang Maha Mengetahui, Maha Bijaksana".
(Q.S Al-Baqarah 32)

Sesungguhnya bersama kesulitan ada kemudahan.
Maka apabila kamu telah selesai (dari suatu urusan),
tetaplah bekerja keras (untuk urusan yang lain).
(Q.S Al-Insyirah 6-7)

Musuh yang paling berbahaya di atas dunia ini adalah penakut dan bimbang. Teman yang paling setia, hanyalah keberanian dan keyakinan yang teguh.
(@Andrew Jackson)

STOP UNDERESTIMATING YOURSELF
~Berehentilah meremehkan diri kamu sendiri~

Skripsi ini kupersembahkan untuk:

- ALLAH SWT, terimakasih telah memberiku kebahagiaan
- Papa dan Mamaku tersayang, atas segala doa, dukungan, serta kasih sayang yang melimpah.
- Kakak dan adikku tersayang, *Jazakumullahu khoiron katsiro*
- Dosen-dosenku yang senantiasa membimbing
- Pihak2 yang belum tersebut disini...makasih
- Almamaterku yang ku banggakan.

ABSTRAK

Nur Ariska, 2017. Analisis *Cluster* dengan Metode *Ensemble ROCK* untuk Data Berskala Campuran Katergorik dan Numerik (Kasus: Mahasiswa Aktif Program Studi Statistika FMIPA UNM). Program Studi Statistika, Fakultas Matematika dan Ilmu Pengetahuan Alam, Universitas Negeri Makassar (dibimbing oleh Muhammad Nusrang dan Sudarmin).

Analisis *cluster* merupakan suatu teknik data mining yang digunakan untuk mengelompokan data berdasarkan kemiripan atribut dari data objek. Salah satu permasalahan yang sering ditemui dalam analisis *cluster* yaitu data yang berskala campuran kategorik dan numerik. Salah satu algoritma yang digunakan untuk memproses data campuran adalah *algCEBMDC (Cluster Ensemble Based Mixed Data Clustering)*. Tahap *clustering* untuk data campuran menggunakan metode *ensemble ROCK (Robust Clustering using linKs)* dilakukan dengan menggabungkan output *clustering* dari data berskala kategorik dan numerik. Metode yang digunakan untuk data kategorik adalah metode *ROCK* dan metode yang digunakan untuk data numerik adalah metode *AGNES (Hierarchical Aggomerative Nesting)*. Adapun metode *clustering* terbaik ditentukan berdasarkan kriteria rasio antara simpangan baku dalam kelompok (S_W) dan simpangan baku antar kelompok (S_B) terkecil. Berdasarkan 107 objek pengamatan, metode *ensemble ROCK* dengan nilai θ sebesar 0,25 menghasilkan dua *cluster* dengan nilai rasio sebesar 0,21 berdasarkan gabungan dari hasil output metode *ROCK* dan metode *AGNES*. Karakteristik hasil *cluster* metode *ensemble ROCK* yang diperoleh menjelaskan bahwa nilai rata-rata IPK yang tinggi terdapat pada *cluster* dua.

Kata kunci: Data Mining, analisis *cluster*, *cluster ensemble algCEBMDC*

ABSTRACT

Nur Ariska, 2017. Cluster Analysis with ROCK Ensemble Methods for Clustering Mixed Categorical and Numerical Dataset (Case: Student Active Study Program Statistics FMIPA UNM). Depatement of Statistics, Faculty of Mathematics and Natural Science. State University of Makassar (supervised by Muhammad Nusrang dan Sudarmin).

Cluster analysis is a data mining technique used to categorize data based on similarity attributes of object data. One of the problems often encountered in clustering analysis is a numerical and categorical dataset. One of the algorithms used to process mixed data is algCEBMDC (Cluster Ensemble Based Mixed Data Clustering). The grouping stage for mixed data uses the ensemble ROCK (Robust Clustering using linKs) method performed by combining grouping outputs from categorical and numerical data. The method used for categorical data is the ROCK method and the method used for numerical data is the AGNES (Hierarchical Aggomerative Nesting) method. Best clustering method is determined by the smallest rasio of standard deviation in groups (S_W) and standard deviation between groups (S_B). Based on 107 observation objects, by using the ensemble ROCK method with values of θ is 0,25 produces two groups of data with ratio value of 0,21, based on a combination of ROCK method output and AGNES method. Characteristics of the cluster of ROCK ensemble methods obtained explained that a high average IPK score is found in cluster two.

Keywords: Data Mining, cluster analysis, cluster ensemble algCEBMDC

KATA PENGANTAR

Syukur Alhamdulillah Robbil Aalamiin, penulis panjatkan kehadirat Allah SWT, yang telah memberi rahmat dan hidayah-nya kepada penulis sehingga dapat menyelesaikan skripsi ini sebagai tugas akhir. Shalawat dan salam semoga tercurah kepada Rasulullah Muhammad SAW, keluarga beliau, para sahabatnya dan seluruh ummatnya yang tetap istiqamah pada ajaran islam.

Skripsi dengan judul **Analisis *Cluster* dengan Metode *Ensemble Rock* untuk Data Berskala Campuran Kategorik dan Numerik (Kasus: Mahasiswa Aktif Program Studi Statistika FMIPA UNM)**. Penulisan ini disusun untuk memenuhi salah satu persyaratan akademik guna memperoleh gelar Sarjana Statistika Fakultas Matematika dan Ilmu Pengetahuan Alam Universitas Negeri Makassar.

Dalam menyusun skripsi ini, penulis mendapatkan sedikit hambatan dan kesulitan yang dialami. Terbatasnya kemampuan, pengetahuan, dan wawasan menjadi hambatan besar dalam penyusunan skripsi ini. Namun berkat kerja keras dari semua pihak, pada akhirnya penulis dapat menyelesaikan dengan semaksimal mungkin. Saran dan kritik yang membangun penulis diharapkan dapat memberikan manfaat bagi peningkatan penulis di masa yang akan datang. Maka melalui pengantar ini penulis menghaturkan terima kasih yang sebesar-besarnya kepada dosen pembimbing yakni bapak Drs. Muhammad Nusrang, M.Si., dan bapak Sudarmin, S.Si., M.Si yang telah berkenan memberikan waktu luang,

arahan, bimbingan serta dengan penuh kesabaran meneliti setiap kata demi kata dalam skripsi ini. Serta kepada dosen penguji yakni bapak Prof. H. M. Arif Tiro, M.Pd., M.Sc., Ph.D, dan bapak Adiatma, S.Pd., M.Si yang telah memberikan masukan dan saran-saran yang membangun dalam penyelesaian skripsi ini. Penulis juga mengucapkan terima kasih kepada seluruh rekan-rekan di kampus yang telah meluangkan waktunya untuk membantu dan mengarahkan penulis, dan kepada teman-teman seperjuangan angkatan 2013 Statistika FMIPA UNM yang telah memberikan dukungan dan bantuan selama mengikuti pendidikan di Kampus Orange.

Penulis menghaturkan pula ucapan terima kasih yang sebesar-besarnya terutama kepada:

1. Bapak Rektor Universitas Negeri Makassar.
2. Bapak Dekan Fakultas Matematika dan Ilmu Pengetahuan Alam Universitas Negeri Makassar yang telah memberikan kelancaran pelayanan dalam urusan akademik.
3. Bapak Ketua Program Studi Statistika Fakultas Matematika dan Ilmu Pengetahuan Alam Universitas Negeri Makassar yang telah mendidik dan memberi motivasi kepada penulis selama dalam proses perkuliahan.
4. Bapak/Ibu Dosen-Dosen Statistika yang telah mendidik, dan memberikan ilmu kepada penulis selama menempuh jenjang pendidikan.

Terwujudnya skrpsi ini adalah berkat do'a, dan restu keluarga tercinta. Oleh karena itu, penulis menghanturkan terima kasih tak terhingga kepada kedua orang tua tercinta, Ayahanda Anas dan Ibunda Hasni yang telah mendidik,

mencurahkan perhatian, kasih sayang, dan do'anya demi kesuksesan dan kebaikan penulis serta bantuan moril maupun material mulai dari ananda lahir hingga menyelesaikan studi sarjana Statistika Fakultas Matematika dan Ilmu Pengetahuan Alam Universitas Negeri Makassar. Semoga apa yang telah beliau berikan kepada penulis menjadi kebaikan dan cahaya penerang kehidupan dunia dan akhirat. Demikian juga buat saudara-saudara tercinta Wahyu Ekafrian, Riski, Nugie Nugraha, dan Aidil Firah atas segala kasih sayang, perhatian dan dukungan yang diberikan kepada penulis selama menempuh pendidikan.

Semoga yang telah penulis sebutkan di atas mendapat imbalan bernilai pahala di sisi Allah SWT, *Aamiin Allahumma Aamiin.* Dengan segala kerendahan hati penulis menyadari sepenuhnya bahwa skripsi ini masih sangat jauh dari kesempurnaan. Oleh karena itu, penulis menerima kritik dan saran yang bersifat membangun. Semoga penulisan skripsi ini dapat bermanfaat bagi pembaca dan pihak yang terkait.

Makassar, Desember 2017

Penulis

Nur Ariska

DAFTAR ISI

DAFTAR TABEL

DAFTAR GAMBAR

DAFTAR LAMPIRAN

BAB I

PENDAHULUAN

A. Latar Belakang

Analisis *cluster* merupakan suatu teknik data mining yang digunakan untuk mengelompokan data berdasarkan kemiripan atribut dari data objek (Rahayu, 2013). Data mining merupakan suatu proses untuk menemukan informasi yang berguna di dalam data dengan ukuran besar secara otomatis (Tan, Steinbach, & Kumar, 2006). Data mining juga merupakan bagian integral dari *Knowledge Discovery in Databases* (KDD), dimana KDD memiliki beberapa proses mulai dari pengumpulan data sampai pada proses mendapatkan informasi. Adapun tujuan utama analisis *cluster* adalah untuk mengelompokan objek-objek pengamatan menjadi beberapa kelompok berdasarkan karakteristik yang dimiliki. Pada umumnya, algoritma analisis *cluster* dikembangkan hanya untuk memproses salah satu tipe data kategorik atau numerik. Permasalahan yang sering di dapat dalam analisis *cluster* adalah jenis data yang berskala campuran kategorik dan numerik. Dewangan, Sharma, & Akasapu (2010) menyatakan bahwa metode yang seringkali dilakukan untuk mengelompokan data yang berskala campuran adalah dengan mentransformasi data kategorik menjadi data numerik dan sebaliknya. Akan tetapi metode tersebut mempunyai kelemahan dalam menentukan transformasi yang tepat agar tidak kehilangan banyak informasi dari original datanya. Berdasarkan kelemahan *clustering* dengan metode transformasi tersebut, maka dikembangkan sebuah metode *clustering ensemble* untuk data berskala

campuran. *Cluster ensemble* adalah suatu metode yang digunakan untuk menjalankan beberapa algoritma *clustering* yang berbeda, untuk mendapatkan bagian yang sama dari data yang bertujuan untuk menyatukan hasil dari hasil-hasil *clustering* individual (Hee, Xu, & Deng, 2002).

Pada umumnya algoritma *clustering* hanya digunakan untuk memproses salah satu tipe data numerik atau kategorik saja. Tidak banyak algoritma *clustering* yang dikembangkan untuk memproses data dengan tipe campuran. Salah satu metode yang dapat digunakan adalah *algCEBMDC (Cluster Ensemble Based Mixed Data Clustering)* yang merupakan suatu algoritma *clustering* dengan pendekatan *cluster ensemble*.

Dalam penelitian ini, *clustering* data numerik dilakukan dengan metode *Algoritma Hierarchical Agglomerative Nesting* (*AGNES*) sedangkan *clustering* data kategorik dilakukan dengan metode *RObust Clustering using linKs* (*ROCK*). Setelah kedua *cluster* dari data numerik dan kategorik terbentuk, selanjutnya *cluster-cluster* yang dihasilkan oleh kedua algoritma digabungkan dan dipandang sebagai data baru dengan tipe kategorik, kemudian diproses dengan menggunakan algoritma *clustering* data kategorik untuk mendapatkan hasil akhir. Algoritma tersebut yang dikatakan dengan *algCEBMDC.*

Adapun data yang digunakan dalam penelitian ini adalah data mining yang merupakan suatu proses untuk menemukan informasi yang menarik dan tersembunyi dari suatu kumpulan data yang berukuran besar yang tersimpan dalam suatu basis data, data warehouse atau tempat penyimpanan data lainnya yaitu data kemahasiswaan Universitas Negeri Makassar khususnya Program Studi

Statistika. Salah satu alasan menggunakan data kemahasiswaan karena dalam data kemahasiswaan biasanya sering tersimpan informasi yang sangat penting tentang Mahasiswa, antara lain tentang demografi dan prestasi akademik mereka sehingga informasi tersebut dapat digunakan oleh pihak institusi untuk menyusun dan mengembangkan program secara lebih tepat, efektif dan efisien (Saxena, Khare, & Garg, 2002). Metode penelitian ini mengikuti alur kerja data mining dan *algCEBMDC*.

B. Rumusan Masalah

Analisis *cluster* merupakan suatu teknik data mining yang digunakan untuk mengelompokan data berdasarkan kemiripan atribut dari data objek. Data mining merupakan suatu proses untuk menemukan informasi yang berguna di dalam data dengan ukuran besar. Data mining mempunyai tipe data yang berbeda-beda. Pada umumnya algoritma *cluster* dikembangkan hanya untuk memproses salah satu tipe data kategorik atau numerik. Adapun permasalahan yang sering di dapat dalam analisis *cluster* adalah jenis data yang berskala campuran kategorik dan numerik. Metode yang seringkali dilakukan untuk mengelompokan data yang berskala campuran adalah dengan mentransformasi data kategorik menjadi data numerik dan sebaliknya. Selain pengelompokan dengan metode transformasi tersebut, dikembangkan sebuah metode *clustering ensemble* untuk data campuran. Salah satu algoritma untuk memproses data campuran adalah *algCEBMDC*.

C. Pertanyaan Penelitian

1. Bagaimana hasil *cluster* yang terbentuk menggunakan metode *ensemble ROCK* untuk data berskala campuran kategorik dan numerik?
2. Bagaimana karakteristik hasil *cluster* yang terbentuk menggunakan metode *ensemble ROCK?*

D. Tujuan Penelitian

1. Untuk mengetahui hasil *cluster* yang terbentuk menggunakan metode *ensemble ROCK* untuk data berskala campuran kategorik dan numerik?
2. Untuk mengetahui karakteristik dari hasil *cluster* yang terbentuk menggunakan metode *ensemble ROCK*?

E. Manfaat Penelitian

1. Manfaat teoritis

Penelitian ini diharapkan dapat menambah wawasan keilmuan mengenai analisis *cluster* dengan pendekatan *algCEBMDC* untuk data berskala campuran kategorik dan numerik

2. Manfaat Praktis

Hasil penelitian ini diharapkan dapat memberikan informasi bagi Universitas Negeri Makassar khususnya Program Studi Statistika, serta untuk membantu pengambilan kesimpulan secara umum berdasarkan hasil analisis

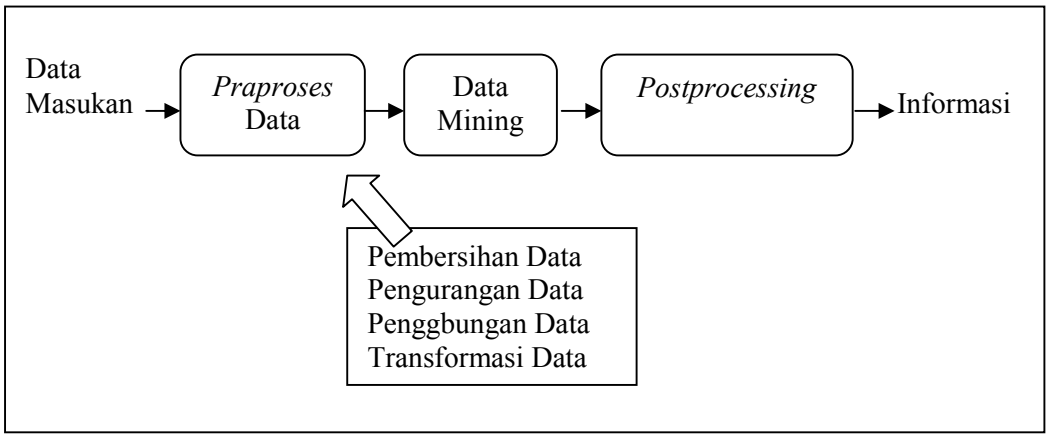
BAB II

KAJIAN PUSTAKA

A. Tinjauan Pustaka

1. Data Mining

Menurut Tan, Steinbach, & Kumar (2006), data mining merupakan suatu proses untuk menemukan informasi yang menarik dan tersembunyi dari suatu kumpulan data yang berukuran besar yang tersimpan dalam suatu basis data, data warehouse atau tempat penyimpanan data lainnya. Teknik-teknik data mining yang digunakan bertugas untuk menemukan pola baru dan bermakna di dalam basis data yang mungkin masih belum diketahui. Data mining juga merupakan bagian integral dari *Knowledge Discovery in Databases* (KDD). Keseluruhan proses KDD dari mulai data sampai menjadi informasi ditunjukkan oleh Gambar 2.1 sebagai berikut.



Gambar 2.1 Proses dari KDD (Tan, Steinbach, & Kumar, 2006)

Berdasarkan gambar tersebut, data masukan mengalami 3 proses sebelum menjadi hasil yang berupa informasi yaitu praproses data, data mining, dan *postprocessing*. Praproses bertujuan untuk mentransformasi data ke dalam format sesuai dengan kebutuhan. Tahapan praproses antara lain adalah pembersihan data untuk membuang data-data yang tidak digunakan dan data duplikat, pengurangan data, penggabungan data, dan transformasi atau normalisasi data. *Postprocessing* bertujuan untuk membantu pengguna dalam memahami informasi. Kualitas informasi yang dihasilkan oleh proses KDD sangat dipengaruhi oleh kualitas data, pengetahuan tentang data, dan teknik pengolahan data yang akan digunakan.

a. Tipe data

Data adalah komponen dasar dalam proses data mining yang merupakan fakta yang diolah menjadi suatu informasi. Setiap data terdiri dari kumpulan data objek/data observasi. Karakteristik dari data objek digambarkan dengan beberapa atribut, dimana setiap atribut memiliki nilai dengan tipe yang berbeda-beda. Secara umum terdapat dua tipe data, yaitu data kategorik dan data numerik. Data kategorik merupakan suatu data dengan peubah kualitatif yang dihasilkan dari pengklasifikasian atau penggolongan suatu data (data atribut) sedangkan data numerik adalah suatu data kuantitatif dimana atribut yang dimilikinya bertipe numerik. Agresti (2006), menyatakan bahwa data kategorik memiliki skala pengukuran yang terdiri atas satu set kategorik.

b. Praproses data

Praproses data dilakukan karena data awal cenderung untuk tidak bersih, tidak lengkap dan tidak konsisten. Praproses data bertujuan untuk meningkatkan kualitas data sehingga diharapkan dapat membantu meningkatkan akurasi, efektifitas, dan efisiensi dari suatu proses data mining. Praproses data merupakan langkah yang sangat penting dalam proses KDD karena kualitas hasil akhir suatu proses data mining sangat dipengaruhi oleh kualitas data. Praproses data juga bertujuan untuk mentranformasi data input ke dalam format sesuai dengan kebutuhan. Pembersihan data, pengurangan data, penggabungan data, dan transformasi data merupakan bagian dari praproses data (Han & Kamber, 2001).

1) Pembersihan data

Pembersihan data dilakukan karena data penelitian seringkali memiliki *record* dengan nilai atribut yang tidak lengkap, nilai kosong, tidak konsisten, dan *noisy*. Data yang memiliki atribut dengan nilai tidak lengkap atau kosong dapat diatasi dengan beberapa cara yaitu menghapus data tersebut, isi atribut kosong dengan rata-rata nilai atribut atau isi atribut kosong dengan nilai atribut yang paling sering muncul (Han & Kamber, 2001). Nilai tidak konsisten adalah nilai yang berada diluar kesepakatan. Data *noisy* adalah kesalahan tidak berpola atau perbedaan yang terjadi pada peubah yang diukur (Tan, Steinbach, & Kumar, 2006).

2) Pengurangan data

Pengurangan data biasanya dikaitkan dengan data yang sangat besar yang merupakan suatu usaha yang digunakan untuk mengurangi ukuran data dengan

tujuan untuk memperoleh data dengan volume yang relatif kecil tetapi dapat mewakili kondisi data asli. Memproses data hasil pengurangan seharusnya jauh lebih efisien dibanding dengan memproses data asli tetapi mendapatkan hasil yang relatif sama. Seleksi atribut dan seleksi *record* merupakan sebagian dari teknik pengurangan data.

a) Seleksi atribut

Data yang akan dianalisis bisa jadi memiliki atribut dengan jumlah yang cukup banyak tetapi sesungguhnya sebagian dari atribut tersebut tidak relevan dengan kebutuhan penelitian. Sebagai contoh, jika akan dilakukan *clustering* terhadap data Mahasiswa untuk menemukan karakteristik Mahasiswa yang berkaitan dengan Indeks Prestasi Akademik, maka atribut seperti Nama, Alamat, atau Nomor Telepon termasuk atribut yang tidak relevan dengan kebutuhan penelitian. Jika atribut tersebut diikutsertakan dalam proses *clustering*, maka selain memperlambat proses, juga akan mendapatkan hasil yang kurang berkualitas. Seleksi atribut adalah suatu usaha untuk mengurangi ukuran data dengan cara menghapus atribut yang tidak relevan dengan kebutuhan penelitian (Han & Kamber, 2006).

b) Seleksi *record*

Secara umum, karakteristik data mining adalah menganalisis data dengan ukuran yang sangat besar berdasarkan sampel dari data tersebut. Sampel digunakan untuk memberikan informasi terkait dengan keseluruhan data. Kualitas dari informasi yang dihasilkan tergantung dari data objek yang dipilih sebagai

sampel. Seleksi *record* adalah suatu usaha untuk mendapatkan data sampel yang representatif dengan data asli.

c) Penggabungan data

Pada proses data mining seringkali dibutuhkan suatu proses penggabungan data. Penggabungan dilakukan karena data yang akan dianalisis berasal dari beberapa sumber. Sumber tersebut dapat berupa *multiple databases*, data *cubes*, atau *flat file*.

d) Transformasi data

Secara prinsip, data kategori dapat ditransformasi/dikonversi ke dalam bilangan numerik, dimana satu bilangan numerik mewakili satu nilai kategori. Atribut kategori yang demikian disebut dengan "*dummy variable*" (Kandardzic, 2011). Dalam suatu data numerik kadang-kadang terdapat atribut yang memiliki nilai dengan rentang yang sangat berbeda dengan atribut lain atau dengan kata lain memiliki satuan yang berbeda. Untuk beberapa algoritma data mining, kondisi demikian dapat mengacaukan hasil perhitungan *proximity* (Tan, Steinbach, & Akasapu, 2006). Atribut dengan rentang nilai besar menjadi sangat dominan, dan akan mempengaruhi hasil secara tidak proporsional. Oleh karenanya, perlu dilakukan standarisasi terhadap semua atribut sehingga setiap atribut memiliki kontribusi secara proporsional terhadap hasil akhir suatu proses data mining

2. Analisis *Cluster*

Analisis *cluster* merupakan suatu metode multivariat yang bertujuan untuk mengelompokan sampel subyek atas dasar satu set peubah yang diukur menjadi

beberapa kelompok yang berbeda sehingga subyek yang sama ditempatkan dalam kelompok yang sama (Cornish, 2007). Menurut Simamora (2005), analisis *cluster* merupakan suatu teknik analisis statistik yang ditujukan untuk menempatkan sekumpulan objek ke dalam dua atau lebih grup berdasarkan kesamaan-kesamaan objek atas dasar berbagai karakteristik. Menurut Han & Kamber (2001), analisis *cluster* adalah suatu teknik data mining untuk mengelompokan himpunan objek (dataset) ke dalam beberapa *cluster* hanya berdasarkan kemiripan karakteristik dari atribut yang dimiliki oleh data objek sedemikian sehingga data objek yang berada di dalam *cluster* yang sama memiliki kemiripan satu sama lain tetapi tidak mirip dengan data objek yang berada dalam *cluster* yang berbeda. Hasil analisis *cluster* dipengaruhi oleh objek yang dikelompokkan, peubah yang diamati, ukuran kemiripan atau ketidakmiripan yang digunakan, skala ukuran yang digunakan, serta metode *clustering* yang digunakan.

a. Ukuran kemiripan

Ukuran kemiripan digunakan untuk mencari pasangan objek yang mirip dalam data. Kemiripan antar pasangan objek x dan y dinyatakan dengan $si\ (x,y)$. $si\ (x,y)$ akan bernilai besar jika x dan y merupakan pasangan objek yang mirip, sebaliknya $si\ (x,y)$ akan bernilai kecil jika x dan y merupakan pasangan objek yang tidak mirip.

Untuk setiap pasangan objek x dan y, berlaku 3 kondisi berikut (Kandardzic, 2011):

1) $0 \leq si\ (x,y) \leq 1$, kemiripan bernilai 0 dan 1.

2) $si\ \ (x,x) = 1$, setiap objek mirip dengan dirinya sendiri.

3) $si\ \ (x,y) = si\ \ (y,x)$, kemiripan bersifat simetri.

b. Ukuran ketidakmiripan

Ukuran ketidakmiripan digunakan untuk mencari jarak antara pasangan objek di dalam data. Jarak antara pasangan objek x dan y dinyatakan dengan $d(x,y).d(x,y)$ akan bernilai besar jika x dan y merupakan pasangan objek yang tidak mirip, sebaliknya $d(x,y)$ akan bernilai kecil jika x dan y merupakan pasangan objek yang mirip. Untuk setiap objek x dan y berlaku kondisi berikut (Han & Kamber, 2001):

1) $d(x,y) \geq 0$, jarak merupakan bilangan non-negatif.

2) $d(x,x) = 0$, jarak suatu objek dengan dirinya sendiri = 0.

3) $\mathrm{d(x,y)} = \mathrm{d(y,x)}$, jarak bersifat simetri.

Semakin besar nilai ukuran ketidakmiripan antara dua objek maka semakin besar pula perbedaan antara kedua objek tersebut, sehingga makin cenderung untuk tidak berada dalam kelompok yang sama (Johnson & Wichern, 2007).

3. Metode *Clustering*

Dalam analisis *cluster*, tahap pengelompokkan dibedakan menurut jenis data yang dimiliki. Pada umumnya analisis *cluster* terfokus pada data numerik, akan tetapi terdapat kasus dengan data kategorik bahkan terdapat kasus dengan campuran data kategorik dan numerik. Analisis *cluster* pada data kategorik tidak dapat diperlakukan seperti pada data numerik. Hal tersebut dikarenakan sifat khusus data kategorik, sehingga *clustering* data kategorik menjadi lebih rumit

dibandingkan *clustering* untuk data numerik (Hair, Black, Babin, & Anderson, 2010).

a. *Clustering* data kategorik

Clustering data kategorik dilakukan dengan menggunakan ukuran kemiripan atau jarak untuk data berskala kategorik kemudian dapat dilakukan *clustering* dengan menggunakan metode hirarki maupun *non*-hirarki. Metode *clustering* hirarki dan *non*-hirarki dinilai tidak tepat digunakan pada data kategorik sehingga dikembangkan metode *ROCK* untuk *clustering* data kategorik tersebut (Guha, Rastogi, & Shim, 1999).

Metode *clustering* yang digunakan untuk tipe data kategorik adalah algoritma *ROCK*. *ROCK* pertama kali diperkenalkan oleh Guha, Rastogi, & Shim pada tahun 1999. Metode *ROCK* menggunakan konsep *link* sebagai ukuran kemiripan untuk membentuk *cluster*-nya. Metode *ROCK* dapat menangani *outlier* dengan cukup efektif. Pemangkasan *outlier* memungkinkan untuk membuang yang tidak ada tetangga, sehingga titik tersebut tidak berpartisipasi dalam pengelompokan. Namun dalam beberapa situasi, *outlier* dapat hadir sebagai *cluster-cluster* yang kecil (Guha, Rastogi, & Shim, 1999).

Clustering untuk data kategorik dengan algoritma *ROCK* dilakukan dengan tiga langkah. Adapun langkahnya yaitu sebagai berikut:

1. menghitung *similaritas* menggunakan rumus *Jaccard coefficient* (Rahayu, 2009). Ukuran kemiripan antara pasangan objek ke$-i$ dan objek ke$-j$ dihitung dengan rumusan yang didefinisikan pada persamaan 2.1.

$$si\ \ (X_{i,}X_{j,}) = \frac{|X_i \cap X_j|}{|X_i \cup X_j|}\ \ , X_i \neq X_j \qquad (2.1)$$

dimana:

$i = 1, 2, 3, \dots, n\ \ d \quad j = 1, 2, 3, \dots, n$

X_i = himpunan penngamatan ke-i dengan $X_i = \left\{x_{1i}, x_{2i}, x_{3i}, \dots, x_{m_k\ \ \ \ \ i}\right\}$

X_j = himpunan penngamatan ke-j dengan $X_j = \left\{x_{1j}, x_{2j}, x_{3j}, \dots, x_{m_k\ \ \ \ \ j}\right\}$

$|X|$ = bilangan kardinal atau jumlah anggota dari himpunan X.

2. Langkah kedua adalah menentukan tetangga. Pengamatan dinyatakan sebagai tetangga jika nilai $si\ \ (X_{i,}X_{j,}) \geq \theta$.

3. Langkah terakhir adalah menghitung *link* antar objek pengamatan. Besarnya *link* dipengaruhi oleh nilai *threshold* (θ) yang merupakan parameter yang ditentukan oleh pengguna yang dapat digunakan untuk mengontrol seberapa dekat hubungan antara objek. Besarnya nilai θ yang diinputkan adalah $0 < \theta < 1$.

Metode *ROCK* menggunakan informasi tentang *link* sebagai ukuran kemiripan antar objek. Jika terdapat objek pengamatan $X_{i,}X_{j,}$, dan $X_{k,}$ dimana $X_{i,}$ tetangga dari X_j, dan X_j tetangga dari X_k maka dikatakan X_i memiliki *link* dengan X_k walaupun X_i bukan tetangga dari X_k. Cara untuk menghitung *link* untuk semua kemungkinan pasangan dari n objek dapat menggunakan matriks $\boldsymbol{A}$. Matriks $\boldsymbol{A}$ merupakan matriks berukuran $n\,x\,n$ yang bernilai 1 jika X_i dan X_j dinyatakan mirip (tetangga) dan bernilai 0 dan jika X_i dan X_j tidak mirip (bukan tetangga). Jumlah *link* antar pasangan X_i dan X_j diperoleh dari hasil kali antara baris ke $X_i i$ dan kolom ke X_j dari matriks **A**. Jika *link* antara X_i dan X_j semakin

besar maka semakin besar pula kemungkinan X_i dan $X_j$berada dalam satu kelompok yang sama.

Adapun metode Penggabungan *cluster* yang digunakan yaitu algoritma *ROCK* yang didasarkan atas ukuran kebaikan (*goodness measure*) antar kelompok dengan rumusan pada persamaan 2.2. *Goodness measure* adalah persamaan yang digunakan untuk menghitung jumlah *link* dibagi dengan kemungkinan *link* yang terbentuk berdasarkan ukuran kelompoknya (Tyagi & Sharma, 2012).

$$g(C_i, C_j) = \frac{li \quad [C_i, C_j]}{(n_i + n_j)^{1+2f(\theta)} - n_i{}^{1+2f(\theta)} - n_j{}^{1+2f(\theta)}} \tag{2.2}$$

dengan $li \quad [C_i, C_j] = \sum_{X_i \square C_i,\ X_j \square C_j,} li \quad (X_i, X_j)$ yang menyatakan jumlah *link* dari semua kemungkinan pasangan objek yang ada dalam C_i dan C_j , serta n_i dan n_j masing-masng menyatakan jumlah anggota dalam kelompok ke- i dan j, sedangkan $f(\theta) = \frac{1-\theta}{1+\theta}$.

b. *Clustreing* data numerik

Clustering data numerik dilakukan berdasarkan ukuran ketidakmiripan atau jarak untuk data numerik. Hasil *clustering* disajikan dalam bentuk *dendrogram* (diagram pohon) yang memungkinkan penelusuran objek-objek yang diamati menjadi lebih mudah dan informatif. Metode *clustering* yang digunakan untuk tipe data numerik adalah algoritma *AGNES*. *AGNES* pertama kali diperkenalkan oleh Kaufmann dan Rousseeuw pada tahun 1990. *AGNES* merupakan algoritma *agglomerative hierarchical clustering* yang cukup popular yang berproses pada data numerik (Han & Kamber, 2001).

Menurut Rencher (2002), dalam setiap langkah pendekatan metode hirarki *agglomerative*, observasi atau kelompok pengamatan tergabung dalam kelompok lain. Algoritma *AGNES* dimulai dengan menghitung matriks jarak antar objek, setiap objek berfungsi sebagai *cluster*, kemudian secara bertahap menggabungkan setiap pasangan *cluster* terdekat berdasarkan ukuran jarak dan metode penggabungan yang digunakan sampai semua *cluster* tergabung dalam satu *cluster*.

Pada peubah yang memiliki jenis skala data numerik maka jarak yang dapat digunakan adalah jarak *euclidean*. Jarak *euclidean* digunakan dalam mengukur jumlah kuadrat perbedaan nilai pada masing-masing peubah.

$$d_{ij} = \sqrt{\sum_{k=1}^{p} \left(X_{ik} - X_{jk}\right)^2} \qquad (2.3)$$

dimana:

d_{ij} = jarak antara objek ke-i dan objek ke-j

p = jumlah peubah *cluster*

X_{ik} = data dari subyek ke- i pada peubah ke- k

X_{jk} = data dari subyek ke- j pada peubah ke- k

Adapun metode penggabungan yang digunakan yaitu metode *single linkage, complete linkage,* dan *average linkage*. Metode penggabungan adalah suatu ukuran kuantitatif yang digunakan oleh algoritma *clustering hierarchical agglomerative* untuk menggabungkan dua *cluster* C_x dan C_y yang dianggap mirip/dekat berdasarkan ukuran jarak kedua *cluster.* Didefinisikan $d(C_x, C_y)$

yang menyatakan jarak antara *cluster* C_x dan C_y, n_x dan n_y masing-masing menyatakan jumlah anggota klaster C_x dan C_y dan $d(x,y)$ menyatakan jarak antara objek x dan y, dimana x dan y masing-masing merupakan anggota *cluster* C_x dan C_y. Metode penggabungan yang sering digunakan yaitu sebagai berikut:

1) Metode *single linkage*

Metode *single linkage* juga biasa di sebut dengan metode tetangga terdekat yang merupakan kesamaan antara *cluster* sebagai jarak yang terpendek dari objek apapun dalam satu *cluster* untuk setiap objek yang lain. Metode ini menggunakan prinsip jarak minimum. Dimulai dengan mencari dua objek yang memiliki jarak terdekat. Keduanya membentuk *cluster* yang pertama. Pada langkah selanjutnya, terdapat dua kemungkinan yaitu objek ketiga akan bergabung dengan *cluster* yang telah dibentuk atau dua objek lain akan membentuk *cluster* baru. Proses ini akan berlanjut sampai akhirnya terbentuk *cluster* tunggal. Pada metode ini, jarak antar *cluster* didefinisikan sebagai jarak terdekat antar anggotanya. Jarak antara *cluster* C_x dan C_y dihitung berdasarkan jarak terdekat antara dua objek dalam *cluster* C_x dan C_y .

Adapun persamaan metode *single linkage* adalah sebagai berikut:

$$d(C_x, C_y) = m \quad d(x,y) \tag{2.4}$$

x □ C_x

y □ C_y

2) Metode *complete linkage*

Metode *complete linkage* juga biasa disebut dengan metode tetangga terjauh yang merupakan kebalikan dari pendekatan yang digunakan pada *single*

linkage. Prinsip jarak yang digunakan adalah jarak terjauh antar objek. Jika dua objek terpisah oleh jarak yang jauh, maka kedua objek tersebut akan digabung menjadi satu *cluster,* demikian seterusnya. Pada metode ini, jarak antar *cluster* didefinisikan sebagai jarak terjauh antar anggotanya. Jarak antara klaster C_x dan C_y dihitung berdasarkan jarak terdekat antara dua objek dalam *cluster* C_x dan C_y.

Adapun persamaan metode *complete linkage* adalah sebagai berikut:

$$d(C_x, C_y) = m \quad d(x, y) \tag{2.5}$$

$x \in C_x$

$y \in C_y$

3) Metode *average linkage*

Metode *avarage linkage* adalah metode *clustering* dengan prinsip jarak rata-rata antar setiap pasangan objek yang mungkin pada satu *cluster* dengan seluruh objek pada *cluster* lain. Pada metode ini, Jarak antara *cluster* C_x dan C_y dihitung berdasarkan rata-rata jarak antara semua kemungkinan pasangan objek dalam *cluster* C_x dan C_y.

Adapun persamaan metode *average linkage* adalah sebagai berikut:

$$d(C_x, C_y) = \frac{1}{n_x n_y} \sum_{x=1}^{n_x} \sum_{y=1}^{n_y} d \tag{2.6}$$

c. *Clustering* data campuran

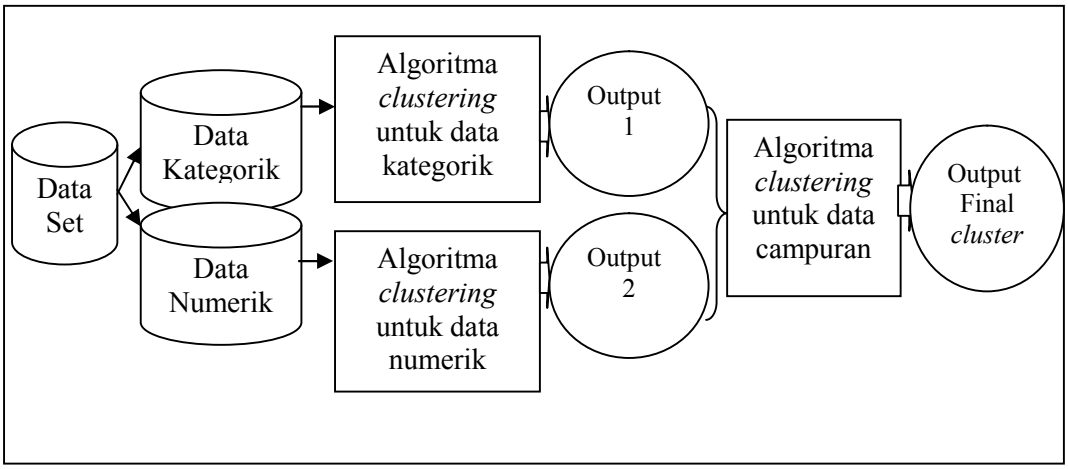
Jika diperhatikan dari tipe data yang akan dianalisa, algoritma *clustering* dibedakan ke dalam tiga jenis yaitu algoritma *clustering* yang digunakan untuk menganalisis data kategorik, algoritma *clustering* yang digunakan untuk

menganalisis data numerik, dan algoritma *clustering* yang digunakan untuk menganalisis data campuran (kategorik dan numerik). Pada umumnya algoritma *clustering* hanya digunakan untuk memproses salah satu tipe data numerik atau kategorik saja. Tidak banyak algoritma *clustering* yang dikembangkan untuk memproses data dengan tipe campuran. Salah satunya adalah *algCEBMDC* yang merupakan suatu algoritma *clustering* dengan pendekatan *cluster ensemble. Cluster ensemble* adalah suatu metode yang digunakan untuk menjalankan beberapa algoritma *clustering* yang berbeda, untuk mendapatkan bagian yang sama dari data, bertujuan untuk menyatukan hasil dari hasil-hasil *clustering* individual (Hee, Xu, & Deng, 2002).

Clustering bertujuan untuk membentuk kelompok serta mendapatkan pola yang menarik dari suatu data. Secara umum output yang dihasilkan oleh suatu algoritma *clustering* menempatkan setiap data objek ke dalam satu *cluster* tertentu. Jika dua objek berada dalam *cluster* yang sama maka kedua objek tersebut dianggap sama. Sebaliknya jika dua objek berada dalam *cluster* yang berbeda maka kedua objek dianggap berbeda. Jelas bahwa *cluster* yang dihasilkan oleh suatu algoritma *clustering* tidak dapat diurutkan sebagaimana mengurutkan bilangan real. Dengan kata lain bahwa *cluster-cluster* tersebut dapat dipandang sebagai data kategori karena output dari masing-masing algoritma *cluster* merupakan data kategori, maka masalah *cluster ensemble* dapat dipandang sebagai masalah dari *clustering* data kategori. Hasil dari masing-masing algoritma *clustering* dapat digabungkan menjadi data yang baru dengan tipe kategori (Hee, Xu, & Deng, 2002).

***algCEBMDC (Cluster Ensemble Based Mixed Data Clustering*)**

Algoritma *algCEBMDC* dikembangkan untuk menyelesaikan masalah yang berkaitan dengan *clustering* data dengan tipe campuran (kategorik dan numerik). Pertama, data asli yang bertipe campuran dipisah menjadi dua yaitu data dengan tipe kategorik dan data dengan tipe numerik. Selanjutnya, kedua data tersebut diproses secara terpisah dengan menggunakan algoritma *clustering* yang sesuai dengan tipe masing-masing data. Terakhir, *cluster-cluster* yang dihasilkan oleh kedua algoritma digabungkan dan dipandang sebagai data baru dengan tipe kategorik, kemudian diproses dengan menggunakan algoritma *clustering* data kategorik untuk mendapatkan hasil akhir (Hee, Xu, & Deng, 2002). Langkah dari *algCEBMDC* ditunjukkan oleh Gambar 2.2 berikut:



Gambar 2.2 Langkah dari *algCEBMDC*

Adapun algoritma AlgCEBMDC menurut Hee, Xu, & Deng (2002) adalah sebagai berikut:

1. pisahkan data menjadi data kategorik dan data numerik
2. lakukan *clustering* terhadap data kategorik dengan menggunakan algoritma *clustering* untuk data kategorik yaitu metode *ROCK*

3. lakukan *clustering* terhadap data numerik dengan menggunakan algoritma *clustering* untuk data numerik yaitu metode *AGNES*
4. gabungkan output dari kedua algoritma tersebut menjadi data kategorik
5. gunakan *ROCK* lagi untuk melakukan *clustering* terhadap data kategorik.

4. Kinerja Hasil *Clustering*

Pengukuran kinerja hasil *clustering* merupakan langkah untuk mengetahui validitas suatu *cluster*. *Cluster* yang baik akan memiliki kehomogenan yang tinggi antar anggota dalam kelompok dan keheterogenan yang tinggi antar kelompok (Hair, Black, Babin, & Anderson, 2010). Adapun kinerja hasil *clustering* untuk peubah dengan skala data numerik berbeda dengan kinerja hasil *clustering* untuk peubah dengan skala data kategorik.

a. Skala data numerik

Kinerja hasil pengelompokan untuk skala data numerik terdiri dari dua uji validasi, yaitu validasi ukuran dan validasi metode.

1) Validasi ukuran

Validasi ukuran yang digunakan dalam pemilihan jumlah *cluster* optimum adalah ukuran *index dunn*. *Index dunn* merupakan salah satu pengukuran validitas *cluster* yang diajukan oleh *J.C.Dunn.* Menurut Satato, Khotimah, & Muhammad (2015), validitas *cluster* berlandaskan pada fakta bahwa *cluster* yang terpisah biasanya memiliki jarak antar *cluster* yang besar dan jarak dalam *cluster* yang kecil. *Indeks dunn* tidak memiliki suatu rentang nilai, untuk mencari *indeks dunn*

terbaik dapat dilihat dari nilai terbesar yang dihasilkan (Dewanti, 2013). Adapun rumus *index dunn* yaitu sebagai berikut:

$$D\ (c) = m_{\ 1\le i\le n}\left\{ m_{\ 1\le j\le n, i\ne j}\left\{ \frac{d(C_i, C_j)}{m_{\ 1\le k\le n}(d'(C_k))} \right\} \right\} \tag{2.7}$$

dimana

$d(C_i, C_j)$ = jarak antara *cluster* C_i dan C_j

$d'(C_k)$ = jarak dalam *cluster* C_k

Nilai terbesar dari *DI* diambil sebagai jumlah optimum *cluster* (Bolshakova & Azuaje, 2001).

2) Validasi metode

Kinerja hasil *clustering* untuk peubah dengan skala data numerik dapat diketahui dari rasio nilai S_W dan S_B. Menurut Bunkers & James (1996), kinerja hasil pengelompokan dengan menggunakan nilai rata-rata peubah, simpangan baku di dalam kelompok atau *within* (S_W) dan simpangan baku antar kelompok atau *between* (S_B) dapat dirumuskan seperti pada persamaan (2.8) dan (2.10) berikut :

$$S_W = \frac{1}{C}\sum_{C=1}^{c} S_c \tag{2.8}$$

dimana:

C = banyaknya *cluster* yang terbentuk

S_c = simpangan baku *cluster* ke-*c.*

Jika diberikan *cluster* c_k, dimana $k = 1, \dots, p$, dan setiap *cluster* memiliki anggota x_i, dimana $i = 1, \dots, n$ dan n adalah jumlah anggota dari setiap *cluster*, dan $\bar{x_k}$ adalah rata-rata dari *cluster* k maka untuk mencari nilai simpangan baku ke- k (s_c) digunakan rumus berikut :

$$s_c = \sqrt{\frac{1}{n-1}\sum_{K=1}^{K}(x_i - \bar{x_k})^2} \tag{2.9}$$

$$S_B = \left[\frac{1}{C-1}\sum_{C=1}^{C}(\bar{x_c} - \bar{x})^2\right]^{1/2} \tag{2.10}$$

dimana:

C = banyaknya *cluster* yang terbentuk

$\bar{x_c}$ = rata-rata *cluster* ke-*c.*

$\bar{x}$ = rata-rata keseluruhan *cluster*

Kinerja suatu metode *clustering* semakin baik, jika semakin kecil nilai rasio antara S_W dan S_B. Hal ini berarti bahwa terdapat homogenitas maksimum dalam *cluster* dan heterogenitas maksimum antar *cluster*.

b. Skala data kategorik

Kinerja hasil *clustering* untuk peubah dengan skala data kategorik adalah dengan menggunakan tabel kontingensi yang ekuivalen dengan melakukan ANOVA *(Analysis of Varianve)*. Menurut Alvionita (2017), ukuran keragaman untuk data kategorik dikembangkan oleh Light dan Nargolin (1971), Okada (1999) serta Kader dan Perry (2007). Jika terdapat sebanyak n pengamatan dengan n_k merupakan jumlah pengamatan dengan kategori ke- k dimana k =

$1, 2, 3, \ldots K$ dan $\sum_{k=1}^{K} n_k = n$. Selanjutnya, n_k merupakan jumlah pengamatan dengan kategori ke-*k* dan kelompok ke-*c,* dimana $c = 1, 2, 3, \ldots C$ dengan C adalah jumlah kelompok yang terbentuk, sehingga $n_{.c} = \sum_{k=1}^{K} n_k$ merupakan jumlah pengamatan pada kelompok ke- c dan $n_{k.} = \sum_{c=1}^{C} n_k$ merupakan jumlah pengamatan pada kategori ke-*k.* Total jumlah pengamatan dapat dituliskan menjadi $n = \sum_{c=1}^{C} n_{.c} = \sum_{k=1}^{K} n_{k.} = \sum_{k=1}^{K} \sum_{c=1}^{C}$.

Jumlah kuadrat total (SST) untuk sebuuah peubah dengan data kategorik dapat dirumuskan seperti persamaan (2.11). untuk total jumlah kuadrat dalam kelompok (SSW) dirumuskan dalam persamaan (2.12), serta jumlah kuadrat antar kelompok (SSB) dapat dirumuskan seperti pada persamaan (2.13). (Alvionita, 2017)

$$S. \quad = \frac{n}{2} - \frac{1}{2n} \sum_{k=1}^{K} n_k^2 \tag{2.11}$$

$$SS \quad = \sum_{c=1}^{C} \left(\frac{n_{.c}}{2} - \frac{1}{2n_{.c}} \sum_{k}^{K} n_k^2 \right) = \frac{n}{2} - \frac{1}{2} \sum_{c=1}^{C} \frac{1}{n_{.c}} \sum_{k}^{K} n_k^2 \tag{2.12}$$

$$S. \quad = \frac{1}{2} \sum_{c=1}^{C} \frac{1}{n_{.c}} \sum_{k}^{K} n_k^2 \ \frac{1}{2n} \sum_{k}^{K} n_{k.}^2 \tag{2.13}$$

*Mean of square total (*MST), *mean of square within (MSW),* dan *mean of square beetwen (MSB).* Dapat dirumuskan sperti pada persamaan (2.14), (2.15), dan (2.16).

$$M \quad = \frac{S.}{(n-1)} \tag{2.14}$$

$$M. \quad = \frac{SS}{(n-C)} \tag{2.15}$$

$$M \quad = \frac{S.}{C-1} \tag{2.16}$$

Simpangan baku dalam kelompok (S_W) dan simpangan baku antar kelompok (S_B) untuk data kategori dapat dirumuskan seoerti pada persamaan (2.17) dan (2.18).

$$S_W \quad = [M. \quad]^{1/2} \tag{2.17}$$

$$S_B \quad = [M \quad]^{1/2} \tag{2.18}$$

Seperti halnya dengan data numerik, kinerja suatu metode pengelompokkan untuk data kategorik semakin baik jika semakn kecil rasio antara S_W dan S_B yang berarti bahwa terdapat homogenitas maksimum dalam *cluster* dan heterogenitas maksimum antar *cluster*.

B. Kerangka Pikir

Analisis *cluster* merupakan suatu teknik data mining yang digunakan untuk mengelompokkan data berdasarkan kemiripan atribut dari data objek. Data mining merupakan suatu proses untuk menemukan informasi yang berguna di dalam data dengan ukuran besar. Data mining mempunyai tipe data yang berbeda-beda. Data yang digunakan dalam penelitian ini adalah data kemahasiswaan Universitas Negeri Makassar khususnya Program Studi Statistika. Salah satu alasan menggunakan data kemahasiswaan karena dalam data kemahasiswaan sering tersimpan informasi yang sangat penting tentang Mahasiswa, antara lain tentang demografi dan prestasi akademik mereka sehingga informasi tersebut dapat digunakan oleh pihak institusi untuk menyusun dan mengembangkan

program secara lebih tepat, efektif dan efisien. Pada umumnya algoritma *cluster* dikembangkan hanya untuk memproses salah satu tipe data kategori atau numerik. Permasalahan yang sering di dapat dalam analisis *cluster* adalah jenis data yang berskala campuran kategorik dan numerik adalah metode *clustering ensembel*. Dalam penelitian ini, pengelompokan data numerik dilakukan dengan metode *AGNES* sedangkan pengelompokan data kategori dilakukan dengan metode *ROCK* Setelah kedua *cluster* dari data kategori dan numerik terbentuk, selanjutnya *cluster-cluster* yang dihasilkan oleh kedua algoritma tersebut digabungkan dan dipandang sebagai data baru dengan tipe kategorik, kemudian diproses dengan menggunakan algoritma *clustering* data kategorik untuk mendapatkan hasil akhir. Algoritma tersebut yang dikatakan dengan *algCEBMDC* untuk mendapatkan final *custer*.

BAB III

METODOLOGI PENELITIAN

A. Sumber Data

Data yang dikumpulkan pada penelitian ini adalah data sekunder yang diperoleh atau dikumpulkan dari basis data Program Studi Statistika Universitas Negeri Makassar angkatan 2013 sampai 2016. Adapun peubah yang digunakan yaitu asal sekolah, status keorganisasian, pekerjaan orangtua, pendidikan terakhir orangtua, IPK, dan SKS.

B. Defenisi Operasional Peubah (DOP)

1. asal sekolah (atribut kategori), berisi kode numerik yang menerangkan pendidikan terakhir sebelum menjadi Mahasiswa Statistika FMIPA UNM. Terdapat tiga kode status pendidikan terakhir dalam data penelitian, yaitu:
   1(SMA), 2(SMK), 3(MA).
2. status keorganisasian (atribut kategori), berisi kode numerik yang menerangkan aktif atau tidak aktif Mahasiswa di keorganisasian dalam kampus maupun luar kampus. Terdapat dua kode untuk status keorganisasian dalam data penelitian, yaitu:
   0 (Tidak Aktif) dan 1(Aktif).
3. pekerjaan orangtua (atribut kategori), berisi kode numerik yang menerangkan pekerjaan orangtua (kepala keluarga). Terdapat beberapa kode pekerjaan campuran. *Cluster ensemble* adalah suatu metode yang digunakan untuk

orang tua dalam data penelitian ini, yaitu: 1(PNS/Pegawai Swasta), 2(Wiraswasta), 3(Petani/nelayan/buruh), 4(lainnya).

4. pendidikan terakhir orangtua (atribut kategori), berisi kode numerik yang menerangkan pendidikan orang tua (kepala keluarga). Terdapat beberapa kode pendidikan terakhir orang tua dalam data penelitian yaitu: 1(S3), 2(S2), 3(S1), 4(DIII), 5(DII)), 6(SMA), 7(SMP), 8(SD), 9(Tidak Tamat SD).
5. IPK (atribut numerik), berisi Indeks Prestasi Kumulatif dari mata kuliah yang berhasil ditempuh dan lulus dengan nilai minimal D. Dalam data penelitian, atribut IPK memiliki rentang 1 s/d 4.
6. SKS (atribut numerik), berisi jumlah Satuan Kredit Semester dari semua matakuliah yang sudah berhasil ditempuh dan lulus dengan nilai minimal D. Dalam data penelitian, atribut SKS memiliki rentang nilai antara 3 s/d 175.

C. Teknik Analisis Data

Metode *ensembel* yang digunakan adalah *algCEBMDC* dimana metode untuk *final cluster* menggunakan metode *ROCK* dengan langkah sebagai berikut:

1. mempersiapkan data
2. melakukan *praproses* data yang dimulai dengan pembersihan data, pengurangan data, pemisahan data, dan terakhir transformasi data.
3. membagi original data yaitu memisahkan peubah yang digunakan menjadi sub-data yang keseluruhan berskala kategori dan keseluruhan berskala numerik.
4. Pengelompokan peubah kategori menggunakan metode *ROCK*

a. melakukan inisialisasi objek sebagai *cluster* dengan anggota tunggal.
b. membentuk similaritas antar objek dengan kriteria menggunakan persamaan 2.1.
c. menentukan *threshold* (θ). Nilai *threshold* (θ) yang digunkan yaitu 0,01, 0,05, 0,10, 0,25, 0,50, 0,75, 0,80, 0,95.
d. menghitung nilai *link* antar pengamatan
e. menghitung nilai *goodness measure* menggunakan persamaan 2.2 sehingga diperoleh *cluster* yang diharapkan.
f. mengulangi langkah (e) dengan nilai θ berbeda.
g. menghitung rasio S_W dan S_B untuk masing-masing nilai θ dengan rumusan pada persamaan 2.17 dan 2.18.
h. membandingkan hasil langkah (g) untuk masing-masing nilai θ dan menentukan jumlah kelompok yang optimum dengan kriteria rasio dengan kriteria rasio S_W dan S_B terkecil.

5. pengelompokan peubah numerik menggunakan metode hirarki *agglomerative*
   a. melakukan inisialisasi objek sebagai kelompok dengan anggota tunggal.
   b. menentukan ukuran ketidakmiripan dengan jarak *euclidean* dengan rumus pada persamaan 2.3 dan membuat matriks jarak berukuran $n\,x\,n$.
   c. menggabungkan kelompok yang memiliki jarak terdekat.
   d. memperbarui matriks jarak dengan metode *single linkage* seperti pada persamaan 2.4.
   e. mengulangi langkah (c) sampai (d) sampai hanya terbentuk 1 *cluster*.

f.. menghitung indeks validitas kelompok menggunakan *Index Duun* seperti pada persamaan 2.7.

g. menentukan kandidat jumlah kelompok yang optimum berdasarkan indeks validitas yang diperoleh pada langkah (f).

h. mengulangi langkah (a) sampai dengan langkah (g) menggunakan metode *complete linkage* seperti pada persamaan 2.5.

i. mengulangi langkah (a) sampai dengan langka (g) menggunakan metode *average linkage* seperti pada persamaan 2.6.

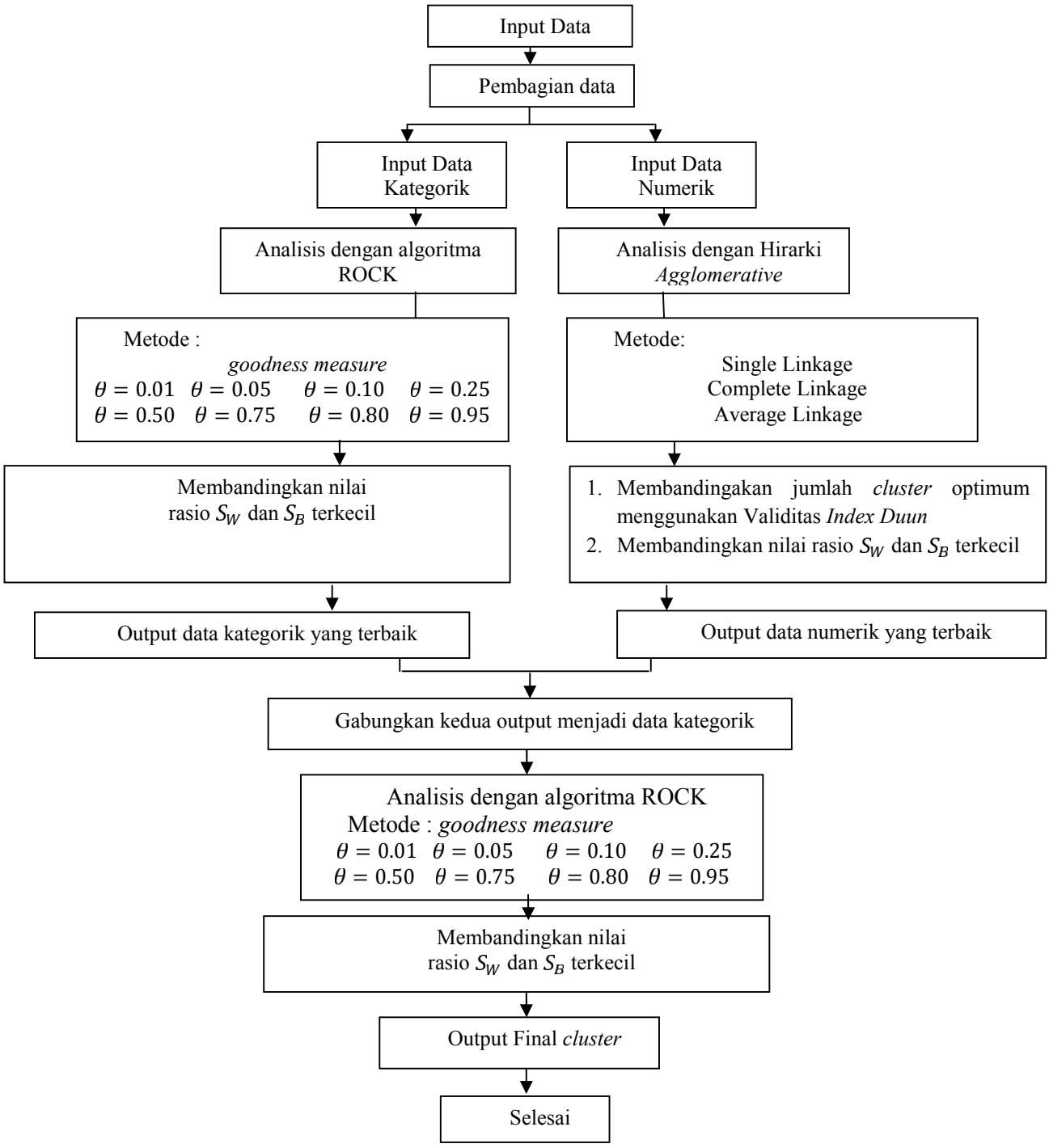
j. menghitung rasio S_W dan S_B dengan rumusan pada persamaan 2.8 dan 2.10 untuk *single linkage, complete linkage,* dan *average linkage*.

k. membandingkan hasil langkah (j) dan menentukan *cluster* terbaik untuk ukuran jarak *euclidean* dengan kriteria rasio S_W dan S_B terkecil.

6. Penggabungan hasil *clustering* (tahapan *ensemble*)

Setelah mendapatkan *cluster* yang optimum hasil metode *ROCK* dan *agglomerative,* tahapan selanjutnya adalah melakukan penggabungan *cluster*. Tahapan ini sama dengan melakukan *clustering* data kategorik menggunakan metode *ROCK* dengan nilai *threshold* (θ) yang digunkan yaitu $0{,}01$, $0{,}05$, $0{,}10$, $0{,}25$, $0{,}50$, $0{,}75$, $0{,}80$ dan $0{,}95$, dimana input untuk tahapan ini adalah *cluster* hasil metode *ROCK* (output 1) dan *cluster* hasil metode *AGNES* (output 2). Output 1 dan output 2 dinyatakan sebagai peubah kategorik yang digunakan untuk menyusun final *cluster*. Final *cluster* yang baik adalah jumlah *cluster* yang memiliki rasio S_W dan S_B terkecil. Nilai rasio dihitung dengan rumusan seperti pada persamaan 2.17 dan 2.18.

Adapun gambar teknik analisis data untuk prodedur *clustering* metode *Ensemble ROCK* sebagai berikut:



Gambar 3.1 Prodedur analisis *clustezr* metode *ensemble ROCK*

BAB IV

HASIL DAN PEMBAHASAN

Pada bab IV, akan dibahas mengenai analisis *cluster* dengan *algCEBMDC* untuk data campuran kategorik dan numerik. Terlebih dahulu, akan dilakukan analisis masing-masing tipe data kategorik dan numerik.

A. Hasil Penelitian

1. Pemisahan data

Untuk kebutuhan penelitian, data Mahasiswa harus dipisah menjadi dua bagian berdasarkan tipe dari atributnya. Struktur data awal dapat dilihat pada Lampiran 1.

Berikut ini adalah adalah data kategorik dan data numerik yang masing-masing disajikan pada Tabel 4.1 dan 4.2.

Tabel 4.1 Contoh Data Kategorik

Asal Sekolah	Status Keorganisasian	Pekerjaan Orangtua	Pendidikan Terakhir Orangtua
1	1	1	2
1	0	2	6
1	0	1	3
1	0	1	3
1	1	1	2

Berdasarkan Tabel 4.1 tersebut, data yang memiliki atribut dengan tipe kategorik diberi nama data kategorik yang memiliki 4 atribut kategorik yaitu asal sekolah, status keorganisasian, pekerjaan orangtua, dan pendidikan terakhir orangtua.

Tabel 4.2 Contoh Data Numerik

IPK	SKS
3,90	155
3,54	154
3,63	153
3,27	147
3,61	153

Berdasarkan Tabel 4.2 tersebut, data yang memiliki atribut dengan tipe numerik diberi nama data numerik, yang memiliki 2 atribut numerik yaitu IPK, dan SKS.

Berikut ini merupakan statistik deskriptif untuk Mahasiswa aktif Program Studi Statistika FMIPA UNM angkatan 2013-2016 sebanyak 107 Mahasiswa dengan 6 peubah. Adapun statistik deskriptifnya yaitu sebagai berikut:

Tabel 4.3 Statistik Deskriptif Peubah Kategorik Asal Sekolah

Peubah Asal Sekolah	Frekuensi	Persentase (%)
SMA	96	89,72
SMK	4	3,74
MA	7	6,54
Jumlah	107	100

Berdasarkan Tabel 4.3 statistik deskriptif untuk peubah kategorik asal sekolah menjelaskan bahwa terdapat 89,72% Mahasiswa berasal dari lulusan SMA, 3,74 % Mahasiswa lulusan SMK serta 6,54% Mahasiswa lulusan MA.

Tabel 4.4 Statistik Deskriptif Peubah Kategorik Status Keorganisasian

Peubah Status Keorganisasian	Frekuensi	Persentase (%)
Aktif	64	59,81
Tidak Aktif	43	40,19
Jumlah	107	100

Berdasarkan Tabel 4.4 statistik deskriptif untuk peubah kategorik status keorganisasian menjelaskan bahwa terdapat 59,81% Mahasiswa yang aktif berorganisasi, selebihnya yaitu 40,19% Mahasiswa yang tidak aktif berorganisasi.

Tabel 4.5 Statistik Deskriptif Peubah Kategorik Pekerjaan Orangtua

Peubah Pekerjaan Orangtua (Kepala Keluarga)	Frekuensi	Persentase (%)
PNS/ Pegawai Swasta	36	33,64
Wiraswasta	26	24,30
Petani/Buruh	31	28,97
Lainnya	14	13,08
Jumlah	107	100

Berdasarkan Tabel 4.5 statistik deskriptif untuk peubah kategorik peubah pendidikan terakhir orangtua (kepala keluarga) menjelaskan bahwa terdapat lulusan terbanyak berasal dari lulusan SMA yaitu sebanyak 31,80%.

Tabel 4.6 Statistik Deskriptif Peubah Kategorik Pendidikan Terakhir Orangtua

Peubah Pendidikan terakhir Orangtua (Kepala keluarga)	frekuensi	Persentase (%)
S3	2	1,87
S2	14	13,10
S1	24	22,40
DIII	3	2,80
DII	1	0,93
SMA	34	31,80
SMP	6	5,61
SD	20	18,70
Tidak Tamat SD	3	2,80
Jumlah	107	100

Berdasarkan Tabel 4.6 statistik deskriptif untuk peubah kategorik pekerjaan orangtua (kepala keluarga) menjelaskan bahwa terdapat 33,64% orangtua Mahasiswa bekerja sebagai PNS/pegawai swasta, 24,30% bekerja sebagai wiraswasta, 28,97% bekerja sebagai petani/buruh serta 13,08% lainnya.

Adapun statistik deskriptif untuk data numerik yaitu sebagai berikut:

Tabel 4.7 Statistik Deskriptif Peubah Numerik

Peubah	N	Min	Max	Mean
IPK	107	2,97	3,93	3,47
SKS	107	40	155	104

Berdasarkan Tabel 4.7 analisis deskriptif untuk peubah numerik tersebut menjelaskan bahwa Mahasiswa aktif Program Studi Statistika FMIPA UNM Angkatan 2013-2016 sebanyak 107 Mahasiswa. Jika ditinjau dari IPK dan SKS menjelaskan bahwa nilai rata-rata IPK 3,47 dimana nilai IPK tertinggi yaitu 3,93 dan IPK terendah yaitu 2,97. IPK tersebut mengikuti SKS yang dilulusi dimana rata-rata SKS yang dilulusi yaitu 104 SKS dengan jumlah SKS tertinggi yaitu 155 dan jumlah SKS terendah yaitu 40.

2. Transformasi Data

Pada Tabel 4.8 ditampilkan contoh hasil *coding* untuk data kategorik yang dimuat pada Tabel 4.1, sedangkan pada Tabel 4.9 ditampilkan contoh hasil standarisasi data numerik yang dimuat pada Tabel 4.2.

Tabel 4.8 Contoh Hasil *Coding* Data Kategorik

Asal Sekolah	Aktif Keorganisasian	Pekerjaan Orangtua	Pendidikan Terakhir Orangtua
10	20	31	41
10	20	31	40
10	20	31	41
10	20	31	40
10	21	31	42

Beberapa atribut dari data kategorik memiliki nilai dengan kode numerik yang sama. Hasil pengkodean tersebut dapat mengacaukan hasil perhitungan

ukuran kemiripan antar objek. Oleh karena itu, dilakukan peng*coding*an terhadap semua atau sebagian atribut yang memiliki kode numerik sama, sedemikian sehingga kode numerik yang dimiliki oleh suatu atribut tidak sama dengan kode numerik yang dimiliki oleh atribut yang lain.

Tabel 4.9 Contoh Hasil Standarisasi Data Numerik

IPK	SKS
1,91	1,16
0,32	1,14
0,72	1,12
-0,88	0,98
0,63	1,12

Data numerik memiliki rentang nilai yang sangat berbeda pada masing-masing atributnya atau dengan kata lain satuan setiap atribut berbeda sehingga dilakukan standarisasi. Sebagai contoh, atribut IPK memiliki rentang nilai antara 0 s/d 4, sedangkan SKS memiliki rentang nilai antara 3 s/d 175. Nilai atribut tersebut memiliki perbedaan yang cukup signifikan yang dapat mengacaukan hasil perhitungan *proximity* antar data objek. Oleh karena itu, perlu dilakukan standarisasi terhadap semua atribut sehingga setiap atribut memiliki kontribusi secara proporsional terhadap hasil akhir suatu proses data mining.

3. *Clustering*

Clustering mahasiswa terdiri dari 3 tahap berdasarkan pemisahan data yaitu data kategorik dan data numerik. Berdasarkan pemisahan data tersebut, sehingga metode untuk masing-masing tipe data akan berbeda pula.

a. *Clustering* data kategorik

Clustering untuk data kategorik menggunakan metode *ROCK*. Tahap pertama yang dilakukan dalam metode *ROCK* adalah menyatakan (inisialisasi) setiap objek pengamatan sebagai suatu *cluster* dengan anggota tunggal. Tahap berikutnya adalah membentuk matriks jarak antar objek pengamatan dengan menggunakan rumus pada persamaan 2.1. Jarak yang diperoleh dari 107 objek pengamatan tersebut dinyatakan dalam matriks si yang berukuran $107\ x\ 107$.

$$si = \begin{bmatrix} 1{,}00 & & & & & & & & \\ 0{,}14 & 1{,}00 & & & & & & & \\ 0{,}33 & 0{,}33 & 1{,}00 & & & & & & \\ 0{,}33 & 0{,}33 & 1{,}00 & 1{,}00 & & & & & \\ \vdots & \vdots & \vdots & \vdots & \ddots & & & & \\ 0{,}14 & 0{,}33 & 0 & 0 & \dots & 1{,}00 & & & \\ 0{,}60 & 0{,}14 & 0{,}60 & 0{,}60 & \cdots & 0{,}14 & 1{,}00 & & \\ 0{,}14 & 0{,}33 & 0{,}33 & 0{,}33 & \cdots & 0{,}00 & 0{,}14 & 1{,}00 & \\ 0{,}60 & 0{,}14 & 0{,}60 & 0{,}60 & \cdots & 0{,}14 & 1{,}00 & 0{,}14 & 1{,}00 \end{bmatrix}$$

Matriks si merupakan matriks yang berisikan jarak dari seluruh kombinasi objek pengamatan dengan diagonal matriks bernilai 1 (jarak objek pengamtan dengan dirinya sendiri). Sebagai contoh, untuk nilai pada baris kedua kolom pertama matriks si tersebut menunjukkan bahwa jarak antara pengamatan pertama dengan pengamatan kedua adalah sebesar $0{,}14$. Setelah diperoleh jarak antara pengamatan, selanjutnya ditentukan nilai θ sebagai batas penentuan tetangga. Informasi mengenai hubungan tetangga antara objek pengmatanan dapat dinyatakan dengan matriks **A**. Matriks **A** merupakan matriks berukuran 107 x 107 yang bernilai 1 jika objek tersebut bertetangga dan bernilai 0 jika objek tersebut tidak bertetangga. Dikatakan bertetangga jika nilai $si > \theta$.

Sebagai contoh, untuk jarak antara pengamatan pertama dengan pengamatan kedua yang bernilai $0{,}14$, maka dengan nilai $\theta = 0{,}25$ dapat

dinyatakan bahwa pengmatan tersebut tidak bertetangga sehingga matiks **A** pada baris kedua kolom pertama bernilai 0.

$$\mathbf{A} = \begin{bmatrix} 1 & & & & & & & & \\ 0 & 1 & & & & & & & \\ 1 & 1 & 1 & & & & & & \\ 1 & 1 & 1 & 1 & & & & & \\ \vdots & \vdots & \vdots & \vdots & \ddots & & & & \\ 0 & 1 & 0 & 0 & \cdots & 1 & & & \\ 1 & 0 & 1 & 1 & \cdots & 0 & 1 & & \\ 0 & 1 & 0 & 1 & \cdots & 0 & 0 & 1 & \\ 1 & 0 & 1 & 1 & \cdots & 0 & 1 & 0 & 1 \end{bmatrix}$$

Setelah diperoleh informasi tetangga antar seluruh kombinasi pengamatan, selanjutnya dilakukan perhitungan jumlah *link* dan *goodness measure*. Perhitungan jumlah *link* dilakukan dengan melakukan perkalian matriks **A** dengan matriks **A** itu sendiri. Perhitungan jumlah *link* tersebut, dinyatakan dalam matiks *link* yang berukuran $107\ x\ 107$.

Dalam penelitian ini digunakan beberapa nilai θ yaitu $\theta = 0{,}01$, $\theta = 0{,}05$, $\theta = 0{,}10$, $\theta = 0{,}25$, $\theta = 0{,}5$, $\theta = 0{,}75$ $\theta = 0{,}80$ dan $\theta = 0{,}95$. Nilai tersebut ditentukan oleh peneliti yang disesuaikan dengan jarak objek pengamatan dan hasil *clustering* yang diharapkan. Hasil yang diharapkan adalah hasil *clustering* dimana semua objek pengamatan tidak berada dalam satu *cluster*, serta tidak terdapat *cluster* dengan anggota tunggal. Hasil *clustering* metode *ROCK* disajikan pada Lampiran 3.

Clustering metode *ROCK* dengan *software R* dapat menghasilkan hasil *clustering* yang berbeda setiap melakukan *running* data. Hal ini dikarenakan adanya perbedaan nilai *goodness measure* yang sama (diambil secara random). Hasil *clustering* terbaik ditentukan dari nilai ratio S_W dan S_B terkecil. Berdasarkan Tabel 4.10, menjelaskan bahwa nilai rasio S_W dan S_B terkecil yaitu $\theta =$

0,01 dengan nilai sebesar 0,85 yang merupakan hasil *cluster* terbaik pada metode *ROCK* untuk data katergorik.

Tabel 4.10 Nilai Ratio Hasil *Cluster* Metode *ROCK*

Nilai θ	Ratio S_W dan S_B
0,01	**0,85**
0,05	0,94
0,10	0,91
0,25	0,96
0,50	0,00
0,75	0,00
0,80	0,00
0,95	0,00

Adapun hasil *cluster* terbaik untuk metode *ROCK* dengan nilai θ sebesar 0,01 yang menghasilkan 2 *cluster* yaitu *cluster* 1 dan *cluster* 2 dengan anggota setiap *cluster* ditunjukkan pada Tabel 4.11. Berikut ini adalah tabel anggota *cluster* untuk metode *ROCK* dengan nilai $\theta = 0{,}01$, yaitu sebagai berikut:

Tabel 4.11 Hasil *Cluster* Metode *ROCK* dengan Nilai $\theta = 0{,}01$

Cluster	Anggota *Cluster*
Cluster 1	Responden 3-4, 7-10, 12-17, 22, 25, 28-30, 32-34, 37-39, 43-44, 47, 53, 56-57, 61-62, 66-68, 73, 76-77, 79, 81, 85-90, 96, 98, 100, 103-104, 106-107.
Cluster 2	Responden 1-2, 5-6, 11, 18-21, 23-24, 26-27, 31, 35-36, 40-42, 45-46, 48-52, 54-55, 58-60, 63-65, 69-72, 74-75, 78, 80, 82-84, 91-95, 97, 99, 101-102, 105.

b. *Clustering* data numerik

Clustering untuk data numerik dilakukan dengan menggunakan metode *AGNES*. Tahap pertama yang dilakukan dalam metode hirarki *agglomerative* adalah menyatakan (inisialisasi) setiap objek pengamatan sebagai suatu kelompok dengan anggota tunggal. Tahap berikutnya adalah membentuk matriks jarak antar

objek pengamatan. Jarak yang digunakan dalam penelitian ini adalah jarak *euclidean* yang dihitung menggunakan persamaan 2.3. Jarak yang diperoleh dari 107 objek pengamatan tersebut dinyatakan dalam matriks $\boldsymbol{d}$ yang berukuran $107\ x\ 107$ **(Lampiran 6).**

$$\boldsymbol{d} = \begin{bmatrix} 1 & & & & & & & & \\ 1{,}59 & 1 & & & & & & & \\ 1{,}19 & 0.39 & 1 & & & & & & \\ 2{,}79 & 1.20 & 1.60 & 1 & & & & & \\ \vdots & \vdots & \vdots & \vdots & \ddots & & & & \\ 3{,}38 & 2{,}63 & 2{,}72 & 2{,}69 & \cdots & 1 & & & \\ 3{,}84 & 2{,}85 & 3{,}02 & 2{,}41 & \cdots & 0{,}66 & 1 & & \\ 3{,}09 & 2{,}57 & 2{,}59 & 2{,}65 & \cdots & 0.4 & 1{,}15 & 1 & \\ 3{,}71 & 2{,}78 & 2{,}93 & 2{,}41 & \cdots & 0.48 & 0{,}17 & 0{,}97 & 1 \end{bmatrix}$$

Matriks $\boldsymbol{d}$ merupakan matriks yang berisikan jarak dari seluruh kombinasi objek pengamatan dengan diagonal matriks bernilai 1 (jarak objek pengamatan dengan dirinya sendiri). Sebagai contoh, untuk nilai pada baris kedua kolom pertama matriks $\boldsymbol{d}$ tersebut menunjukkan bahwa jarak antara objek pertama dengan objek kedua adalah sebesar $1{,}59$. Setelah diperoleh nilai jarak anta objek pengamatan, selanjutnya dilakukan pengabungan kelompok dengan jarak terdekat dan perbaharui matriks jarak menggunakan metode penggabungan dengan beberapa teknik *clustering* yaitu *single linkage*, *complete linkage,* dan *average linkage.* Jumlah *cluster* yang dibentuk berdasarkan dendogram untuk ketiga metode tersebut adalah dua *cluster* sampai lima *cluster*. Setelah diperoleh hasil *clustering*, tahap berikutnya adalah menghitung indeks validitas ukuran jumlah *cluster* optimum menggunakan *index dunn.* Penaksiran jumlah kelompok optimum dilakukan dengan melihat nilai terbesar dari indeks validitas *cluster* tersebut.

Tabel 4.12 Hasil Nilai *Index Dunn* Metode *AGNES*

Jumlah *cluster*	*Single Linkage*	*Complete Linkage*	*Average Linkage*
2 *cluster*	**0,22**	0,07	**0,17**
3 *cluster*	0,17	0,08	0,11
4 *cluster*	0,17	0,09	0,09
5 *cluster*	0,12	**0,14**	0,12

Berdasarkan Tabel 4.13 hasil nilai validitas *index dunn* menunjukkan bahwa jumlah *cluster* optimum yang terbentuk untuk ketiga metode tersebut yaitu 2 *cluster* untuk metode *single linkage,* 5 *cluster* untuk metode *complete linkage* dan 2 *cluster* untuk metode *average linkage*.

Setelah memperoleh jumlah *cluster* optimum, selanjutnya dipilih metode *clustering* terbaik dari ketiga metode tersebut berdasarkan nilai ratio S_W dan S_B terkecil dari masing-masing metode. Nilai ratio yang terbentuk disajikan pada Tabel 4.13.

Tabel 4.13 Nilai Ratio Hasil *Cluster* Metode *AGNES*

	Nilai S_W	Nilai S_B	Ratio
Single linkage	0,46	0,96	0,47
Complete linkage	0,28	0,79	**0,36**
Average linkage	0,59	0,62	0,95

Dengan menggunakan rumus pada persamaan 2.8 dan 2.10, diperoleh metode terbaik yaitu metode *complete linkage* yang memiliki nilai rasio S_W dan S_B terkecil yaitu 0,36 (Tabel 4.17). Hal ini menunjukkan bahwa *clustering* data numerik metode *complete linkage* dengan 5 *cluster* merupakan *clustering* yang tepat untuk metode hirarki *agglomerative*.

Berikut ini adalah tabel anggota *cluster* untuk metode *average linkage.*

Tabel 4.14 Angota *Cluster* Metode *Complete Linkage*

No. *Cluster*	Anggota *Cluster*
Cluster 1	Responden 1, 10, 13, 16, 30, 36, 38-39, 44, 51, 66-67, 70, 73
Cluster 2	Responden 2-6, 8-9, 12, 14-15, 17-19, 22-24, 26, 31-32, 34-35, 37, 40-43, 45-46, 48-50, 52, 31-32, 34-35, 37, 40-43, 45-46, 48-50, 52
Cluster 3	Responden 7, 11, 20-21, 25, 27-29, 33, 47
Cluster 4	Responden 53-58, 60-65, 69, 72, 74-77, 79, 81-82, 85, 88, 90, 92, 106
Cluster 5	Responden 59, 68, 71, 78, 80, 83-84, 86-87, 89, 91, 93-105, 107

c. *Clustering* Data Campuran

Tahapan pertama dalam melakukan analisis *cluster ensemble ROCK* untuk data campuran adalah dengan melakukan *clustering* masing-masing jenis data menggunakan metodenya masing-masing. Hasil *clustering* untuk data numerik yang diperoleh menggunaan metode *ROCK* dinyatakan sebagai *output* 1, serta hasil *clustering* untuk data kategorik yang diperoleh menggunakan metode *AGNES* dinyatakan sebagai *output* 2. Berikutnya kedua hasil *output clustering* tersebut dinyatakan sebagai peubah kategorik (tahap *ensemble*) yang kemudian dilakukan *clustering* menggunakan metode *ROCK*.

Dalam analisis ini digunakan beberapa nilai θ seperti pada *clustering* data kategorik yaitu nilai $\theta = 0{,}01$, $\theta = 0{,}05$, $\theta = 0{,}10$, $\theta = 0{,}25$, $\theta = 0{,}5$, $\theta = 0{,}75$ $\theta = 0{,}80$ dan $\theta = 0{,}95$. Hasil *clustering* metode *ensemble ROCK* disajikan pada Lampiran 11. Hasil *clustering* terbaik ditentukan dari nilai ratio S_W dan S_B terkecil. Nilai ratio S_W dan S_B dapat dilihat pada tabel 4.15, sebagai berikut:

Tabel 4.15 Nilai Ratio Hasil *Cluster* Metode *Ensemble ROCK*

Nilai θ	Nilai Ratio
0,01	0,22
0,05	0,95
0,10	0,64
0,25	**0,21**
0,50	0,61
0,75	0,77
0,80	0,82
0,95	0,82

Tabel 4.15 menunjukkan bahwa nilai rasio terkecil merupakan *clustering* dengan nilai θ sebesar $0{,}25$ dengan ratio S_W dan S_B bernilai $0{,}21$. Nilai tersebut menunjukkan bahwa simpangan baku dalam *cluster* bernilai $0{,}21$ kali dari simpangan baku antar *cluster*. Dengan kata lain, variansi data dalam *cluster* memberikan nilai simpangan lebih kecil dibandingkan variansi antar *cluster*.

Adapun hasil *cluster* terbaik untuk nilai θ sebesar $0{,}25$ tersebut merupakan hasil *running* pertama yang menghasilkan 2 *cluster* yaitu *cluster* 1 dan *cluster* 2 dengan anggota setiap *cluster* ditunjukkan pada Tabel 4.16.

Tabel 4.16 Hasil *Cluster* Metode *Ensemble* ROCK dengan Nilai $\theta = 0.25$

Cluster	Anggota *Cluster*
Cluster 1	Responden 1-2, 5-10, 12-15, 19, 23, 25-26, 32, 34-35, 37, 42, 45, 47-48, 54, 57-58, 60, 66-73, 77-81, 89, 91-93, 97-103.
Cluster 2	Responden 3-4, 11, 16-18, 20-22, 24, 27-31, 33, 36, 38-41, 43-44, 46, 49-53, 55-56, 59, 61-65, 74-76, 82, 84, 86-88,90,94-96,106-107.

Adapun karakteristik hasil *cluster* metode *ensemble ROCK* dapat dilihat pada tabel berikut:

Tabel 4.17 Karakteristik Peubah Numerik Metode *Ensemble ROCK*

	Jumlah Mahasiswa	IPK	SKS
Cluster 1	56	3,46	100
Cluster 2	51	3,48	107

Tabel 4.17 merupakan peubah numerik hasil *cluster* dari metode *ensemble ROCK* yang menghasilkan dua *cluster* yaitu *cluster* 1 dan *cluster* 2.

Tabel 4.18 Karakteristik Peubah Kategori Metode *Ensemble ROCK*

	Kategori	*Cluster* 1	*Cluster* 2
Asal Sekolah	SMA	47,66%	42,05%
Status Keorganisasian	Aktif berorganisas	35,71%	40,05%
Pekerjaan Orangtua	PNS/Pegawai Swasta	14,95%	18,70%
Pendidikan Terakhir Orangtua	SMA	17,75%	14,01%

Berdasarkan tabel 4.21 dan 4.22, menjelaskan bahwa karakteristik hasil *clustering* metode *ensemble ROCK* yang diperoleh adalah sebagai berikut:

a. *Cluster* 1

Cluster 1 merupakan *cluster* yang beranggotakan 56 dari 107 Mahasiswa. Berdasarkan peubah numerik (Tabel 4.21), *cluster* tersebut memiliki nilai rata-rata IPK 3,46 dengan rata-rata SKS yang dilulusi adalah 100 SKS. Berdasarkan peubah kategorik (Tabel 4.22), *cluster* tersebut menjelaskan bahwa terdapat 47,66% Mahasiswa berasal dari lulusan SMA, dan terdapat *35,71*% Mahasiswa yang aktif berorganisasi, sedangkan jika ditinjau dari pekerjaan orangtua dan pendidikan terakhir orangtua menjelaskan bahwa sebanyak 56 Mahasiswa

terdapat 14,95% orangtua Mahasiswa bekerja sebagai PNS/pegawai swasta dan 17,75% pendidikan terkahir orangtua berasal dari lulusan SMA.

b. *Cluster* 2

Cluster 2 merupakan *cluster* yang beranggotakan 51 dari 107 Mahasiswa. Berdasarkan peubah numerik (Tabel 4.21), *cluster* tersebut memiliki nilai rata-rata IPK 3,48 dengan rata-rata SKS yang dilulusi yaitu 107 SKS. Berdasarkan peubah kategorik (Tabel 4.22), *cluster* tersebut menjelaskan bahwa terdapat 42,05% Mahasiswa berasal dari lulusan SMA, dan terdapat 40,05% Mahasiswa yang aktif berorganisasi, sedangkan jika ditinjau dari pekerjaan orangtua dan pendidikan terakhir orangtua menjelaskan bahwa sebanyak 51 Mahasiswa terdapat 18,70% orangtua Mahasiswa bekerja sebagai PNS/pegawai swasta dan terdapat 14,01% pendidikan terkahir orangtua berasal dari lulusan SMA.

B. Pembahasan

1. Karakteristik Responden

Penelitian analisis *cluster ensemble ROCK* untuk data campuran ini melibatkan 107 objek. Adapun objeknya itu merupakan Mahasiswa Program Studi Statistika FMIPA UNM dengan dua jenis skala data yang digunakan yaitu data berskala kategorik dan numerik. Data berskala kategorik diantaranya asal sekolah, status keorganisasian, pekerjaan orangtua dan pendidikan terakhir orangtua. Berdasarkan data tersebut menjelaskan bahwa Mahasiswa Statistika FMIPA UNM dominan lulusan SMA serta aktif berorganisasi dan untuk pekerjaan orangtua dominan bekerja sebagai PNS/Pegawai Swasta dan pendidikan terakhir orangtua

paling banyak SMA. Data bersakala numerik diantaranya IPK dan SKS. Berdasarkan data tersebut menjelaskan bahwa dari 107 Mahasiswa memiliki nilai rata-rata IPK 3,47 dimana nilai IPK tertinggi yaitu 3,93 dan IPK terendah yaitu 2,97. IPK tersebut mengikuti SKS yang dilulusi dimana rata-rata SKS yang dilulusi yaitu 104 SKS dengan jumlah SKS tertinggi yaitu 155 dan jumlah SKS terendah yaitu 40.

2. Karakteristik Hasil *Cluster* Metode *Ensemble ROCK*

Hasil *clustering* untuk data campuran menggunakan metode *ensemble ROCK* dengan nilai θ yang digunakan yaitu $\theta = 0{,}01$, $\theta = 0{,}05$, $\theta = 0{,}10$, $\theta = 0{,}25$, $\theta = 0{,}5$, $\theta = 0{,}75$, $\theta = 0{,}80$ dan $\theta = 0{,}95$ menunjukkan bahwa hasil *cluster* dengan nilai $\theta = 0{,}25$ merupakan nilai θ terbaik berdasarkan nilai ratio S_W dan S_B terkecil yaitu 0,21 yang menghasilkan 2 *cluster* yaitu *cluster* 1 dan *cluster* 2.

Hasil *cluster* 1 berdasarkan peubah numerik menjelaskan bahwa nilai rata-rata IPK pada *cluster* tersebut yaitu 3,46 dengan nilai rata-rata SKS yaitu 100 SKS. Berdasarkan peubah kategorik menjelaskan bahwa rata-rata Mahasiswa yang berasal dari lulusan SMA yaitu 47,66%, dan 35,71% Mahasiswa yang aktif berorganisasi, sedangkan untuk pekerjaan orangtua dan pendidikan terakhir orangtua menjelaskan bahwa terdapat 14,95% orangtua Mahasiswa bekerja sebagai PNS dan 17,75% pendidikan terakhir orangtua dominan berasal dari lulusan SMA.

Hasil *cluster* 2 berdasarkan peubah numerik menjelaskan bahwa nilai rata-rata IPK pada *cluster* tersebut yaitu 3,48 dan nilai rata-rata SKS yaitu 107 SKS.

Berdasarkan peubah kategorik menjelaskan bahwa rata-rata Mahasiswa yang berasal dari lulusan SMA 42,05%, dan 40,05% Mahasiswa yang aktif berorganisasi, sedangkan untuk pekerjaan orangtua dan pendidikan terakhir orangtua menjelaskan bahwa terdapat 18,70% orangtua Mahasiswa bekerja sebagai PNS dan 64,01% dominan pendidikan terakhir orangtua berasal dari lulusan SMA.

BAB V

PENUTUP

A. Kesimpulan

Tujuan dari penelitian ini adalah membentuk *cluster* menggunakan metode *ensemble ROCK* untuk data campuran kategorik dan numerik serta mengetahui karakteristik dari hasil *cluster* yang terbentuk menggunakan *algCEBMDC*. Dari hasil penelitian dapat disimpulkan bahwa:

1. Hasil *clustering* data kategorik menggunakan metode *ROCK* dengan nilai $\theta = 0{,}01$, $\theta = 0{,}05$, $\theta = 0{,}10$, $\theta = 0{,}25$, $\theta = 0{,}5$, $\theta = 0{,}75$, $\theta = 0{,}80$ dan $\theta = 0{,}95$. Berdasarkan nilai ratio S dan S terkecil menunjukkan bahwa nilai $\theta = 0{,}01$ merupakan nilai θ terbaik dalam analisis *cluster* untuk data kategorik.
2. Hasil *clustering* data numerik menggunakan metode *AGNES* menunjukkan bahwa metode terbaik untuk data numerik yaitu metode *average linkage* dengan 5 *cluster* optimum.
3. Hasil *clustering* data campuran kategorik dan numeik menggunakan metode *ensemble ROCK* dengan $\theta = 0{,}01$, $\theta = 0{,}05$, $\theta = 0{,}10$, $\theta = 0{,}25$, $\theta = 0{,}5$, $\theta = 0{,}75$, $\theta = 0{,}80$ dan $\theta = 0{,}95$. menunjukkan bahwa nilai $\theta = 0.25$ merupakan nilai θ terbaik dalam analisis *cluster* untuk data campuran kategorik dan numerik. Hasil *cluster* tersebut menjelaskan bahwa rata-rata nilai IPK yang tingi terdapat pada *cluster* dua.

B. Saran

Adapun saran yang dapat diberikan untuk pengembangan dalam penelitian selanjutnya yaitu sebagai berikut:

1. Pendekatan *clustering* data numerik pada penelitian ini adalah dengan metode hirarki *agglomerative* dengan jarak *euclidean* dan metode yang digunakan yaitu *single linkage, complete linkage* dan *average linkage*, sehingga masih terdapat beberapa metode *clustering* data numerik lain seperti metode ward dan ukuran jarak lain seperti *squared euclidean, mahalanobis, manhattan*, *chebychev*.
2. Pendekatan *clustering* data kategorik pada penelitian ini dalah dengan metode *ROCK*, sehingga dilakukan pengembangan dengan metode pengelompokan data kategorik lain seperti metode *Clustering Categorical Data Using Summaries* (CACTUS).
3. Pendekatan *clustering ensembel* pada penelitian ini adalah dengan algoritma *algCEBMDC,* sehingga dilakukan pengembangan dengan pendekatan lain seperti *Similarity Weight and Filter Method* (SWFM).

DAFTAR PUSTAKA

Agresti, A. (2002). *Categorial data analysis (second ed.).* New York: John Wiley & Sond, Inc.

Alvionita. (2017). *Metode ensemble ROCK dan SWFM untuk pengelompokan data campuran numerik dan kategori pada kasus aksesi jeruki [Thesis]*. Surabaya: Institut Teknologi Sepuluh November.

Bolshakova, N., & Azuaje, F. (2001). *Improving Expression Data Mining through Cluster Validity. Departement of Computer Science.* Ireland: Trinity College Dublin.

Bunkers, M. J. (1996). definition of climate regions in the northern plains using an objective cluster modification technique. *J.Climate* , Vol. 9.

Cornish, R. (2007). Statistics: Cluster Analysis. *Mathematics Learning Support Center.*

Dewangan, R. R., Sharma, L. K., & Akasapu, A. K. (2010). Fuzzy clustering technique for numerical and categorical dataset. *International Journal on Computer Science and Enginering* .

Dewanti. (2013). *Perbandingan Metode Cluster validity pada jenis data numerik dan kategori [Skripsi]*. Bogor: Institut Pertanian Bogor.

Guha, S., Rastogi, R., & Shim, K. (1999). *ROCK : A robust clustering algorithm for categorical attributes*.

Hair, JR.J.F., Black, W. C., Babin, B. J., & Anderson, R. E. (2010). *Multivariate data analysis.* United State of America: Prentice-Hall International,Inc.

Han, J., & Kamber, M. (2001). *Data Mining : Concepts and Techniques.* USA: Academic Press.

Hee, Z., Xu, X. i., & Deng, S. (2002). *Clustering mixed numeric and categorical data: A cluster ensemble approach.* China: Harbin Institute of technology.

Johnson, R.A. & Whinchern, D.W. (2007). *Applied multivariate statistical analysis sixth edition.* Prentice Hall: New Jersey.

Kandardzic, M. (20011). *Data Mining: Concepts, Models, Methods, and Algorithms.* USA : John Wiley & Son, Inc.

Rahayu, D. P. (2013). *Analisis karakteristik kelompok dengan menggunakan pendekatan cluster ensemble [Thesis]*. Banten: Universitas Terbuka.

Rahayu, D. P. (2009). *analisis karakteristik mahasiswa non aktif universitas terbuka dengan pendekatan ensemble* . Bogor: Institut Pertanian Bogor.

Rencher, Alfin C. (2002). *Methods of Multivariate Analysis.* Second Edition. New York: Jhon Wiley & Sons, Inc.

Saxena, a., Khare, P., & Garg, S. (2002). *Application of cluster analysis as a tool to analyse distance education students.* India: Indira Gandhi National Open University.

Simamora, B. (2005). *Analisis multivariat pemasaran edisi pertama.* Jakarta: PT. Gramedia Pustaka Utama.

Satato, B. D., Khotimah, B. K., & Muhammad, A. (2015). Pengelompokan Tingkat Kesehatan Masyarakat menggunakan Shelf Organizing Maps Dengan Cluster Validation Idb dan I-Dunn. *Seminar Nasional Aplikasi Teknologi Informasi.*

Tan, P., Steinbach, M., & Kumar, V. (2006). *Introduction to Data Mining.* USA: Pearson Education,Inc .

Tyagi, A., & Sharma, S. (2012). Implementation of ROCK clustering algorithm for the optimazation of query searching time. *International Journal on Computer Science and Engineering* , Vol 4, No 05.

.

LAMPIRAN

Lampiran 1. Data Peubah Kategorik dan Numerik

a. Data Peubah Numerik

No. Objek	IPK	SKS
1	3,9	155
2	3,54	154
3	3,63	153
4	3,27	147
5	3,61	153
6	3,44	150
7	3,11	147
8	3,55	153
9	3,55	155
10	3,93	155
11	3,21	149
12	3,32	152
13	3,89	155
14	3,63	155
15	3,45	150
16	3,8	153
17	3,28	149
18	3,35	147
19	3,38	147
20	3,14	147
21	3,01	143
22	3,38	147
23	3,49	150
24	3,6	153
25	2,97	154
26	3,35	149
27	3,21	149
28	3,21	149
29	3,18	146
30	3,88	155
31	3,27	151
32	3,55	149
33	3,2	147
34	3,3	147
35	3,27	118
36	3,86	83
37	3,56	140

No. Objek	IPK	SKS
38	3,8	139
39	3,92	139
40	3,42	137
41	3,52	139
42	3,5	141
43	3,47	136
44	3,76	141
45	3,66	141
46	3,53	140
47	3,04	129
48	3,29	134
49	3,52	138
50	3,26	134
51	3,81	139
52	3,57	139
53	3,47	86
54	3,74	88
55	3,44	86
56	3,55	84
57	3,43	86
58	3,53	86
59	3,14	84
60	3,45	86
61	3,62	86
62	3,58	86
63	3,61	86
64	3,63	86
65	3,64	86
66	3,85	88
67	3,91	88
68	3,16	82
69	3,63	88
70	3,81	88
71	3,35	84
72	3,67	88
73	3,81	88
74	3,63	88
75	3,49	86
76	3,64	86
77	3,7	88

No. Objek	IPK	SKS
78	3,26	86
79	3,51	86
80	3,16	82
81	3,55	86
82	3,73	40
83	3,44	40
84	3,38	40
85	3,61	40
86	3,39	40
87	3,46	40
88	3,54	40
89	3,31	40
90	3,63	40
91	3,3	40
92	3,56	40
93	3,28	40
94	3,26	40
95	3,42	40
96	3,24	40
97	3,23	40
98	3,38	40
99	3,21	40
100	3,31	40
101	3,2	40
102	3,04	40
103	3,39	40
104	3,41	40
105	3,26	40
106	3,52	40
107	3,3	40

b. Data Peubah Numerik

No. Objek	Asal sekolah	status keorganisasian	pekerjaan orangtua	pendidikan terakhir orangtua
1	10	21	31	42
2	10	20	32	46
3	10	20	31	43
4	10	20	31	43
5	10	21	31	42
6	10	20	31	42
7	10	20	34	48
8	10	20	32	46
9	10	20	32	46
10	10	20	31	42
11	10	21	32	47
12	10	20	32	48
13	10	21	34	43
14	10	20	32	46
15	10	20	31	43
16	13	20	33	48
17	12	21	33	48
18	10	20	32	46
19	10	21	34	43
20	10	21	31	42
21	10	20	31	42
22	10	20	33	49
23	10	21	33	49
24	10	20	34	46
25	10	21	34	46
26	10	20	34	46
27	12	20	32	46
28	10	20	31	43
29	10	20	31	43
30	10	21	33	47
31	13	21	32	42
32	10	20	33	46
33	10	21	31	43
34	10	20	31	43
35	10	20	32	46
36	10	21	32	46
37	10	20	31	48
38	10	21	34	48

No. Objek	asal sekolah	status keorganisasian	pekerjaan orangtua	pendidikan terakhir orangtua
39	10	21	31	43
40	13	21	32	46
41	10	20	32	46
42	10	21	32	43
43	10	21	33	46
44	10	20	32	48
45	10	20	34	46
46	10	21	31	45
47	10	21	34	43
48	10	20	31	43
49	10	21	31	46
50	10	20	31	42
51	10	21	31	41
52	10	20	31	41
53	10	20	33	48
54	10	21	31	42
55	10	21	34	46
56	10	20	32	47
57	10	21	33	48
58	10	21	33	48
59	10	20	31	42
60	10	21	34	46
61	10	21	34	48
62	10	21	33	46
63	13	21	33	48
64	10	21	33	46
65	10	21	33	48
66	10	21	33	48
67	10	21	32	46
68	10	21	31	42
69	10	21	32	43
70	10	21	32	46
71	10	21	31	44
72	10	20	32	46
73	10	21	33	48
74	10	20	33	46
75	10	21	33	46
76	10	21	32	48
77	10	21	31	42
78	10	21	33	48

No. Objek	asal sekolah	status keorganisasian	pekerjaan orangtua	pendidikan terakhir orangtua
79	10	21	33	49
80	10	21	33	47
81	10	21	31	43
82	10	20	31	42
83	10	21	31	42
84	10	21	33	46
85	13	21	32	43
86	10	20	31	43
87	10	20	33	48
88	10	20	31	43
89	10	21	33	48
90	10	21	33	43
91	10	20	32	46
92	13	21	33	46
93	10	21	33	46
94	10	20	31	43
95	10	21	33	46
96	10	21	31	43
97	12	21	32	46
98	10	21	34	44
99	10	21	31	43
100	10	21	32	47
101	10	21	33	48
102	12	21	33	48
103	10	20	34	46
104	13	20	32	46
105	10	21	31	43
106	10	20	33	47
107	10	21	31	43

Lampiran 2. *Syntax* Metode ROCK untuk Peubah Kategori

```
dk<-
    data.frame(DataMhs$`asalsekolah`,DataMhs$`statuskeorganisasian`,DataMh
    s$`pekerjaan orangtua`,DataMhs$`pendidikan terakhir orangtua`)

x<-dummy.data.frame(dk)


rc.01<-rockCluster(x,n=3,theta = 0.01,debug = FALSE)
rc.05<-rockCluster(x,n=3,theta = 0.05,debug = FALSE)
rc.10<-rockCluster(x,n=3,theta = 0.10,debug = FALSE)
rc.25<-rockCluster(x,n=3,theta = 0.25,debug = FALSE)
rc.50<-rockCluster(x,n=3,theta = 0.50,debug = FALSE)
rc.75<-rockCluster(x,n=3,theta = 0.75,debug = FALSE)
rc.80<-rockCluster(x,n=3,theta = 0.80,debug = FALSE)
rc.95<-rockCluster(x,n=3,theta = 0.95,debug = FALSE)

rf.01<-fitted(rc.01)
rf.05<-fitted(rc.05)
rf.10<-fitted(rc.10)
rf.25<-fitted(rc.25)
rf.50<-fitted(rc.50)
rf.75<-fitted(rc.75)
rf.80<-fitted(rc.80)
rf.95<-fitted(rc.95)

theta.01<-rf.01$cl
theta.05<-rf.05$cl
theta.10<-rf.10$cl
theta.25<-rf.25$cl
theta.50<-rf.50$cl
theta.75<-rf.75$cl
theta.80<-rf.80$cl
theta.95<-rf.95$cl


cluster<-
data.frame(theta.01,theta.05,theta.10,theta.25,theta.50,theta.75,theta.80,theta.95)
hasil1<-data.frame(cluster)
```

Lampiran 3. Output Hasil Metode *ROCK* untuk Peubah Kategorik

	theta.01	theta.05	theta.10	theta.25
1	3	3	3	2
2	3	2	3	2
3	2	2	3	2
4	2	2	2	3
5	3	2	2	2
6	3	3	3	2
7	2	2	2	2
8	2	3	2	3
9	2	3	2	2
10	2	2	3	3
11	3	2	3	3
12	2	2	3	2
13	2	2	2	2
14	2	3	3	3
15	2	2	2	2
16	2	3	3	2
17	2	2	3	2
18	3	2	3	2
19	3	2	3	2
20	3	3	2	2
21	3	2	3	3
22	2	3	2	3
23	3	3	3	3
24	3	3	2	2
25	2	2	2	2
26	3	2	2	3
27	3	2	3	3
28	2	3	3	3
29	2	2	2	2
30	2	3	3	2
31	3	2	2	2
32	2	2	3	3
33	2	2	2	2
34	2	3	2	2
35	3	3	2	2
36	3	2	3	3
37	2	2	3	3
38	2	2	3	3
39	2	2	3	2
40	3	2	2	2
41	3	3	2	3
42	3	3	2	2
43	2	3	3	3
44	2	3	2	3
45	3	2	2	2

	theta.01	theta.05	theta.10	theta.25
46	3	2	3	3
47	2	3	2	2
48	3	2	3	3
49	3	3	3	2
50	3	3	3	3
51	3	2	3	3
52	3	3	2	2
53	2	2	2	3
54	3	2	3	3
55	3	2	2	2
56	2	3	2	2
57	2	2	2	2
58	3	2	3	3
59	3	3	3	2
60	3	3	3	3
61	2	3	3	2
62	2	3	2	3
63	3	2	2	3
64	3	3	2	2
65	3	3	2	2
66	2	3	2	2
67	2	3	2	2
68	2	3	3	3
69	3	2	3	3
70	3	2	2	2
71	3	2	2	2
72	3	2	2	3
73	2	3	3	3
74	3	2	3	3
75	3	2	2	2
76	2	3	3	3
77	2	2	3	3
78	3	2	3	2
79	2	3	2	3
80	3	2	3	3
81	2	3	2	3
82	3	2	2	3
83	3	2	3	3
84	3	3	3	2
85	2	3	3	3
86	2	2	3	3
87	2	2	2	3
88	2	2	2	2
89	2	2	2	2
90	2	2	2	3

```
      theta.01      theta.05      theta.10      theta.25
90           2             2             2             3
91           3             3             3             2
92           3             2             2             3
93           3             2             2             3
94           3             3             2             2
95           3             3             3             2
96           2             3             3             3
97           3             3             2             2
98           2             2             3             3
99           3             3             3             3
100          2             3             3             3
101          3             2             2             3
102          3             2             2             2
103          2             2             3             2
104          2             2             2             3
105          3             2             3             3
106          2             3             2             3
107          2             2             2             3
```

```
    theta.01    freq
1          2      52
2          3      55

    theta.05    freq
1          2      61
2          3      46

    theta.10    freq
1          2      54
2          3      53

    theta.25    freq
1          2      52
2          3      55
```

	theta.50	theta.75	theta.80	theta.95
1	1	1	1	1
2	1	1	1	1
3	1	1	1	1
4	1	1	1	1
5	1	1	1	1
6	1	1	1	1
7	1	1	1	1
8	1	1	1	1
9	1	1	1	1
10	1	1	1	1
11	1	1	1	1
12	1	1	1	1
13	1	1	1	1
14	1	1	1	1
15	1	1	1	1
16	1	1	1	1
17	1	1	1	1
18	1	1	1	1
19	1	1	1	1
20	1	1	1	1
21	1	1	1	1
22	1	1	1	1
23	1	1	1	1
24	1	1	1	1
25	1	1	1	1
26	1	1	1	1
27	1	1	1	1
28	1	1	1	1
29	1	1	1	1
30	1	1	1	1
31	1	1	1	1
32	1	1	1	1
33	1	1	1	1
34	1	1	1	1
35	1	1	1	1
36	1	1	1	1
37	1	1	1	1
38	1	1	1	1
39	1	1	1	1
40	1	1	1	1
41	1	1	1	1
42	1	1	1	1
43	1	1	1	1
44	1	1	1	1
45	1	1	1	1
46	1	1	1	1

	theta.50	theta.75	theta.80	theta.95
47	1	1	1	1
48	1	1	1	1
49	1	1	1	1
50	1	1	1	1
51	1	1	1	1
52	1	1	1	1
53	1	1	1	1
54	1	1	1	1
55	1	1	1	1
56	1	1	1	1
57	1	1	1	1
58	1	1	1	1
59	1	1	1	1
60	1	1	1	1
61	1	1	1	1
62	1	1	1	1
63	1	1	1	1
64	1	1	1	1
65	1	1	1	1
66	1	1	1	1
67	1	1	1	1
68	1	1	1	1
69	1	1	1	1
70	1	1	1	1
71	1	1	1	1
72	1	1	1	1
73	1	1	1	1
74	1	1	1	1
75	1	1	1	1
76	1	1	1	1
77	1	1	1	1
78	1	1	1	1
79	1	1	1	1
80	1	1	1	1
81	1	1	1	1
82	1	1	1	1
83	1	1	1	1
84	1	1	1	1
85	1	1	1	1
86	1	1	1	1
87	1	1	1	1
88	1	1	1	1
89	1	1	1	1
90	1	1	1	1
91	1	1	1	1
92	1	1	1	1

```
      theta.50      theta.75     theta.80      theta.95
93           1             1            1             1
94           1             1            1             1
95           1             1            1             1
96           1             1            1             1
97           1             1            1             1
98           1             1            1             1
99           1             1            1             1
100          1             1            1             1
101          1             1            1             1
102          1             1            1             1
103          1             1            1             1
104          1             1            1             1
105          1             1            1             1
106          1             1            1             1
107          1             1            1             1
```

```
   theta.50    freq
1         1       107

   theta.75    freq
1         1       107

   theta.80    freq
1         1       107

   theta.95    freq
1         1       107
```

Lampiran 4. *Syntax* Metode AGNES untuk Peubah Numerik

```
# DataMhs

    dataNumerik<-data.frame(DataMhs$IPK,DataMhs$SKS)

# Standarisasi Variabel

    StdMhsIPK<-scale(DataMhs$IPK, center = TRUE, scale = TRUE)
    StdMhsSKS<-scale(DataMhs$SKS, center = TRUE, scale = TRUE)
    StdNumerik<-data.frame(StdMhsIPK,StdMhsSKS)
```

METODE *SINGLE LINKAGE*

```
# Ukuran jarak
    d<-dist(StdNumerik, method = "euclidean")

# Analisis Cluster Hirarki metode single linkage
    fit.sin<-hclust(d,method ="single")

# Dendogram
    plot(fit.sin)

#Memotong Dendogram untuk k Cluster  (k=2 sampai k=5)
    single<-cutree(fit.sin, k=k)
    rect.hclust(fit.sin,k=k,border = "red")


# mengitung nilai index Dunn untuk menentukan jumlah cluster optimum
    DataMhsStats<-StdNumerik
    MhsStats<-DataMhsStats[,c("StdMhsIPK","StdMhsSKS")]
    Dist<-dist(MhsStats,method = "euclidean")
    clustobj<-hclust(Dist,method = "single")

  Untuk 2 Cluster (k=2)
    nc<-2
    cluster2<-cutree(clustobj,nc)
    dunn(Dist,cluster2)

  Untuk 3 Cluster (k=3)
    nc<-3
    cluster3<-cutree(clustobj,nc)
    dunn(Dist,cluster3)
```

- Untuk 4 *Cluster* (k=4)

```
nc<-4
cluster4<-cutree(clustobj,nc)
dunn(Dist,cluster4)
```

- Untuk 5 *Cluster* (k=5)

```
nc<-5
cluster5<-cutree(clustobj,nc)
dunn(Dist,cluster5)
```

METODE *COMPLETE LINKAGE*

 # Ukuran jarak

```
d<-dist(StdNumerik, method = "euclidean")
```

 # Analisis *Cluster* Hirarki metode *Complete linkage*

```
fit.com<-hclust(d,method ="complete")
```

#Memotong Dendogram untuk k *Cluster* (k=2 sampai k=10)

```
complete<-cutree(fit.com, k=k)
rect.hclust(fit.com,k=k,border = "red")
```

 # Dendogram

```
plot(fit.com)
```

 # Mengitung nilai *index Dunn* untuk menentukan jumlah *cluster* optimum

```
DataMhsStats<-StdNumerik
MhsStats<-DataMhsStats[,c("StdMhsUmur","StdMhsIPK","StdMhsSKS")]
Dist<-dist(MhsStats,method = "euclidean")
clustobj<-hclust(Dist,method = "complete")
```

- Untuk *2 Cluster* (k=2)

```
nc<-2
cluster2<-cutree(clustobj,nc)
dunn(Dist,cluster2)
```

- Untuk 3 *Cluster* (k=3)

```
nc<-3
cluster3<-cutree(clustobj,nc)
dunn(Dist,cluster3)
```

- Untuk 4 *Cluster* (k=4)

```
nc<-4
cluster4<-cutree(clustobj,nc)
dunn(Dist,cluster4)
```

- Untuk 5 *Cluster* (k=5)

```
nc<-5
cluster5<-cutree(clustobj,nc)
dunn(Dist,cluster5)
```

METODE *AVERAGE LINKAGE*

```
# Ukuran jarak
    d<-dist(StdNumerik, method = "euclidean")

# Analisis Cluster Hirarki metode Average linkage
    fit.ave<-hclust(d,method ="average")

# Dendogram
    plot(fit.ave)

#Memotong Dendogram untuk k Cluster (k=2 sampai k=10)
    average<-cutree(fit.ave, k=k)
    rect.hclust(fit.ave,k=k,border = "red")

# Mengitung nilai index Dunn untuk menentukan jumlah cluster optimum
    DataMhsStats<-StdNumerik
    MhsStats<-DataMhsStats[,c(StdMhsIPK","StdMhsSKS")]
    Dist<-dist(MhsStats,method = "euclidean")
    clustobj<-hclust(Dist,method = "average")
  + Untuk 2 Cluster (k=2)
    nc<-2
    cluster2<-cutree(clustobj,nc)
    dunn(Dist,cluster2)
  + Untuk 3 Cluster (k=3)
    nc<-3
    cluster3<-cutree(clustobj,nc)
    dunn(Dist,cluster3)
  + Untuk 4 Cluster (k=4)
    nc<-4
    cluster4<-cutree(clustobj,nc)
    dunn(Dist,cluster4)
  + Untuk 5 Cluster (k=5)
    nc<-5
    cluster5<-cutree(clustobj,nc)
    dunn(Dist,cluster5)
```

Lampiran 5. Output Hasil Standarisasi Peubah Numerik

	StdMhsIPK	StdMhsSKS
1	1.912189169	1.1638685
2	0.317870407	1.1413284
3	0.716450098	1.1187883
4	-0.877868664	0.9835479
5	0.627876833	1.1187883
6	-0.124995915	1.0511681
7	-1.586454781	0.9835479
8	0.362157040	1.1187883
9	0.362157040	1.1638685
10	2.045049066	1.1638685
11	-1.143588458	1.0286280
12	-0.656435503	1.0962483
13	1.867902537	1.1638685
14	0.716450098	1.1638685
15	-0.080709283	1.0511681
16	1.469322847	1.1187883
17	-0.833582032	1.0286280
18	-0.523575606	0.9835479
19	-0.390715709	0.9835479
20	-1.453594884	0.9835479
21	-2.029321103	0.8933876
22	-0.390715709	0.9835479
23	0.096437246	1.0511681
24	0.583590201	1.1187883
25	-2.206467633	1.1413284
26	-0.523575606	1.0286280
27	-1.143588458	1.0286280
28	-1.143588458	1.0286280
29	-1.276448355	0.9610078
30	1.823615905	1.1638685
31	-0.877868664	1.0737082
32	0.362157040	1.0286280
33	-1.187875090	0.9835479
34	-0.745008767	0.9835479
35	-0.877868664	0.3298856
36	1.735042640	- 0.4590171
37	0.406443672	0.8257673
38	1.469322847	0.8032273
39	2.000762434	0.8032273
40	-0.213569180	0.7581471

	StdMhsIPK	StdMhsSKS
41	0.229297143	0.8032273
42	0.140723878	0.8483074
43	0.007863981	0.7356070
44	1.292176318	0.8483074
45	0.849309995	0.8483074
46	0.273583775	0.8257673
47	-1.896461207	0.5778265
48	-0.789295400	0.6905269
49	0.229297143	0.7806872
50	-0.922155296	0.6905269
51	1.513609479	0.8032273
52	0.450730304	0.8032273
53	0.007863981	-0.3913969
54	1.203603053	-0.3463167
55	-0.124995915	-0.3913969
56	0.362157040	-0.4364770
57	-0.169282548	-0.3913969
58	0.273583775	-0.3913969
59	-1.453594884	-0.4364770
60	-0.080709283	-0.3913969
61	0.672163466	-0.3913969
62	0.495016937	-0.3913969
63	0.627876833	-0.3913969
64	0.716450098	-0.3913969
65	0.760736730	-0.3913969
66	1.690756008	-0.3463167
67	1.956475802	-0.3463167
68	-1.365021619	-0.4815572
69	0.716450098	-0.3463167
70	1.513609479	-0.3463167
71	-0.523575606	-0.4364770
72	0.893596627	-0.3463167
73	1.513609479	-0.3463167
74	0.716450098	-0.3463167
75	0.096437246	-0.3913969
76	0.760736730	-0.3913969
77	1.026456524	-0.3463167
78	-0.922155296	-0.3913969
79	0.185010511	-0.3913969\
80	-1.365021619	-0.4815572
81	0.362157040	-0.3913969
82	1.159316421	-1.4282404
83	-0.124995915	-1.4282404

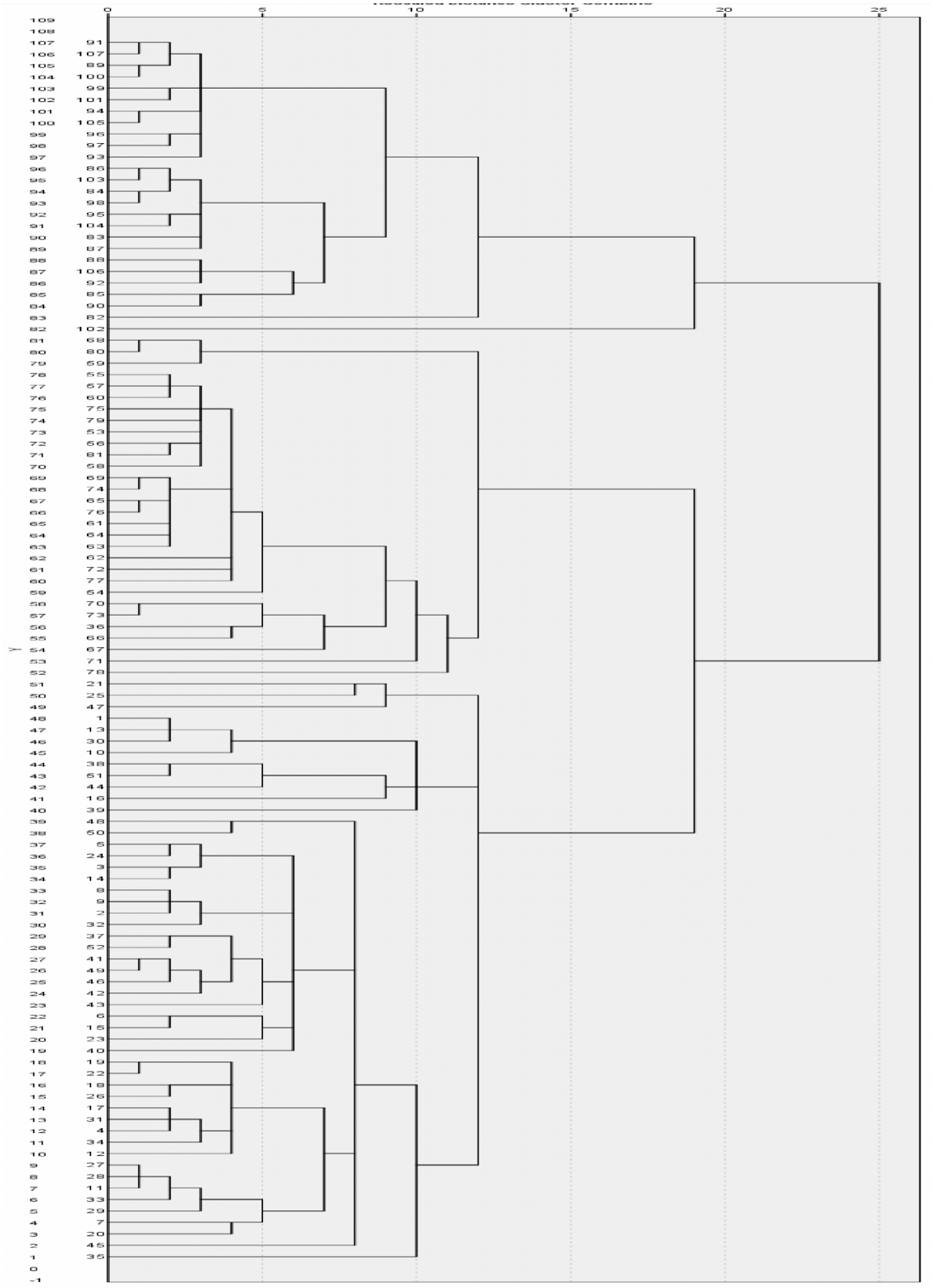
	StdMhsIPK	StdMhsSKS
84	- 0.390715709	-1.4282404
85	0.627876833	-1.4282404
86	- 0.346429077	-1.4282404
87	- 0.036422651	-1.4282404
88	0.317870407	-1.4282404
89	- 0.700722135	-1.4282404
90	0.716450098	-1.4282404
91	- 0.745008767	-1.4282404
92	0.406443672	-1.4282404
93	- 0.833582032	-1.4282404
94	- 0.922155296	-1.4282404
95	- 0.213569180	-1.4282404
96	- 1.010728561	-1.4282404
97	- 1.055015193	-1.4282404
98	- 0.390715709	-1.4282404
99	- 1.143588458	-1.4282404
100	- 0.700722135	-1.4282404
101	- 1.187875090	-1.4282404
102	- 1.896461207	-1.4282404
103	- 0.346429077	-1.4282404
104	- 0.257855812	-1.4282404
105	- 0.922155296	-1.4282404
106	0.229297143	-1.4282404
107	- 0.745008767	-1.4282404

Lampiran 6. Output Hasil Jarak *Euclidean* Metode *AGNES*

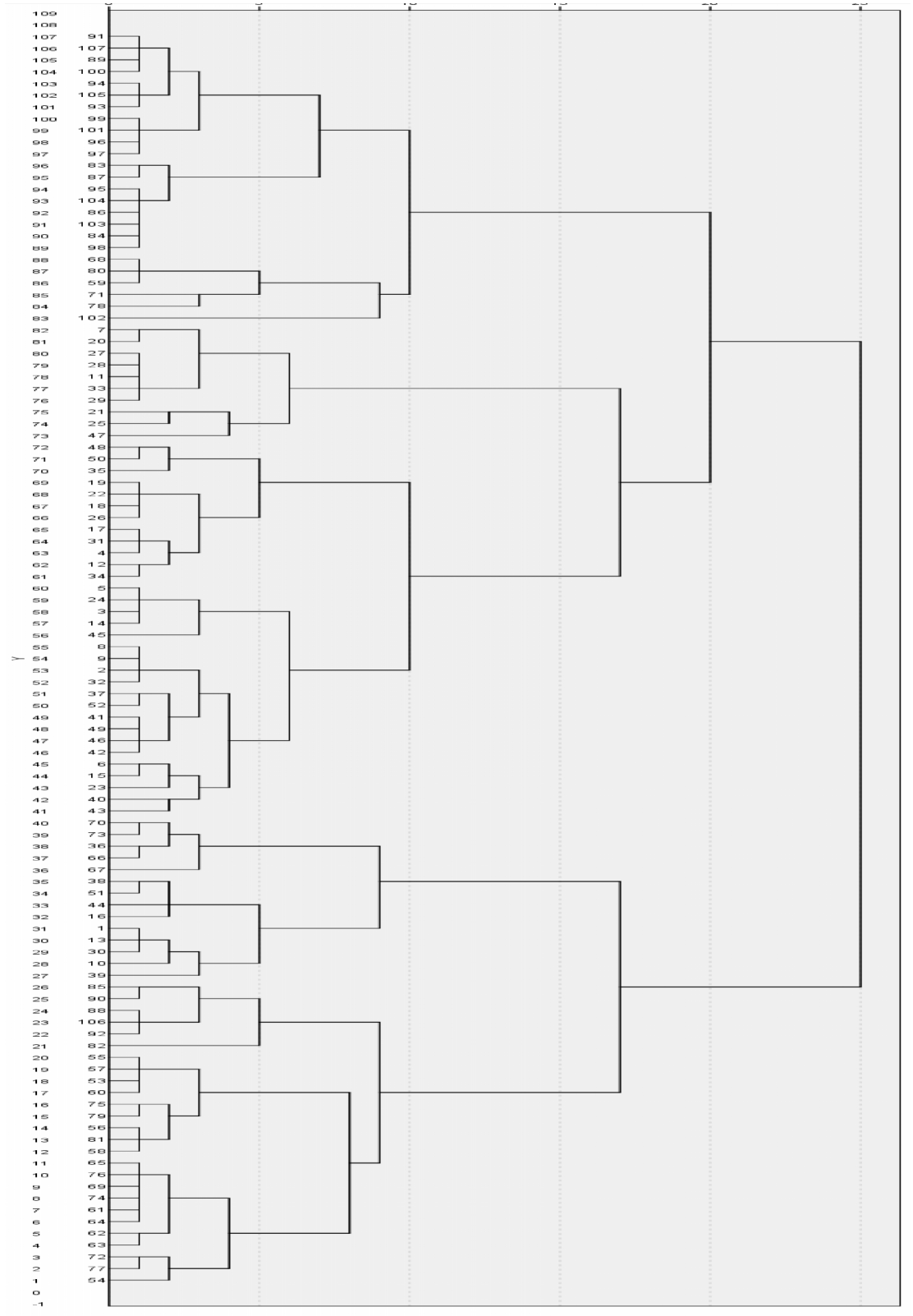
Case	1	2	3	4	5	6	7	8	9	10	11	. . .	100	101	102	103	104	105	106	107
1	0,000	1,594	1,197	2,796	1,285	2,040	3,503	1,551	1,550	0,133	3,059	. . .	3,681	4,041	4,607	3,438	3,381	3,841	3,090	3,712
2	1,594	0,000	0,399	1,206	0,311	0,452	1,911	0,050	0,050	1,727	1,466	. . .	2,764	2,978	3,392	2,654	2,633	2,853	2,571	2,781
3	1,197	0,399	0,000	1,600	0,089	0,844	2,307	0,354	0,357	1,329	1,862	. . .	2,915	3,180	3,649	2,760	2,727	3,029	2,593	2,937
4	2,796	1,206	1,600	0,000	1,512	0,756	0,709	1,247	1,253	2,928	0,270	. . .	2,418	2,432	3,618	2,470	2,490	2,412	2,654	2,415
5	1,285	0,311	0,089	1,512	0,000	0,756	2,218	0,266	0,270	1,418	1,774	. . .	2,873	3,128	3,586	2,727	2,697	2,982	2,578	2,893
6	2,040	0,452	0,844	0,756	0,756	0,000	1,463	0,492	0,500	2,173	1,019	. . .	2,545	2,698	3,047	2,489	2,483	2,604	2,505	2,556
7	3,503	1,911	2,307	0,709	2,218	1,463	0,000	1,953	1,957	3,636	0,445	. . .	2,569	2,445	2,432	2,712	2,754	2,502	3,019	2,554
8	1,551	0,050	0,354	1,247	0,266	0,492	1,953	0,000	0,045	1,683	1,508	. . .	2,760	2,982	3,404	2,644	2,621	3,853	2,550	2,777
9	1,550	0,050	0,357	1,253	0,270	0,500	1,957	0,045	0,000	1,683	1,512	. . .	2,802	3,020	3,438	2,687	2,665	2,893	2.596	2,819
10	0,133	1,727	1,329	2,928	1,418	2,173	3,636	1,683	1,683	0,000	3,192	. . .	3,776	4,144	4,717	3,527	3,467	3,940	3,165	3,808
11	3,059	1,466	1,862	0,270	1,774	1,019	0,445	1,508	1,512	3,192	0,000	. . .	2,496	2,457	2,570	2,583	2,612	2,467	2,814	2,489
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮		⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
100	3,681	2,764	3,229	2,915	2,418	2,873	2,545	2,569	3,760	2,802	3,776	. . .	0,000	0,487	1,196	0,354	0,443	0,221	0,930	0,044
101	4,041	2,978	4,223	3,180	2,432	3,128	2,698	2,445	2,982	3,020	4,144	. . .	0,487	0,000	0,709	0,841	0,930	0,221	1,417	0,443
102	4,607	3,392	5,037	3,649	2,618	3,586	3,047	2,432	3,404	3,438	4,717	. . .	1,196	0,709	0,000	1,550	1,639	0,266	2,126	1,151
103	4,438	2,654	3,458	2,760	2,470	2,727	2,489	2,712	2,644	2,687	3,527	. . .	0,354	0,841	1,550	0,000	0,089	0,576	0,576	0,399
104	3,381	2,633	3,432	2,727	2,490	2,697	2,483	2,754	2,621	2,665	3,467	. . .	0,443	0,930	1,639	0,089	0,000	0,664	0,487	0,487
105	3,841	2,853	4,608	3,029	2,412	2,982	2,604	2,502	2,853	2,893	3,940	. . .	0,221	0,266	0,974	0,576	0,664	0,000	1,151	0,177
106	3,090	2,571	3,327	2,593	2,654	2,578	2,505	3,019	2,550	2,596	3,165	. . .	0,930	1,417	2,126	0,576	0,487	1,151	0,000	0,974
107	3,712	2,781	4,043	2,937	2,415	2,893	2,556	2,554	2,777	2,819	3,808	. . .	0,044	0,443	1,151	0,399	0,847	0,177	0,974	0,000

Lampiran 7. Output Hasil Dendogram Metode *AGNES*

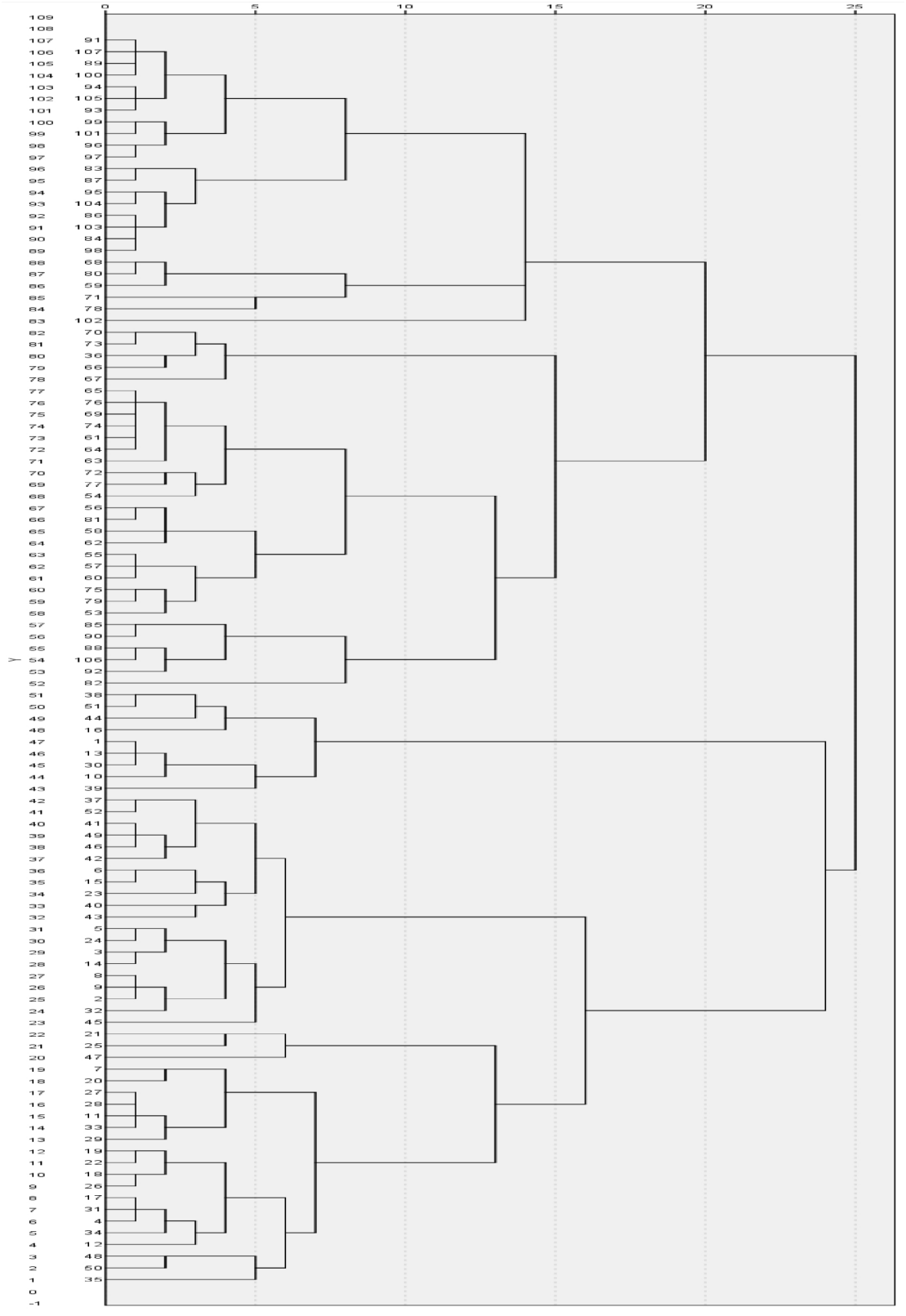
Single Linkage



Complete Linkage



Average Linkage



Lampiran 8. Output Hasil Jumlah *Cluster* Optimum Metode *AGNES*

a. Analisis *cluster* yang dibentuk sama dengan dua (k = 2)

	Single	complete	average
1	1	1	1
2	1	2	1
3	1	2	1
4	1	2	1
5	1	2	1
6	1	2	1
7	1	2	1
8	1	2	1
9	1	2	1
10	1	1	1
11	1	2	1
12	1	2	1
13	1	1	1
14	1	2	1
15	1	2	1
16	1	1	1
17	1	2	1
18	1	2	1
19	1	2	1
20	1	2	1
21	1	2	1
22	1	2	1
23	1	2	1
24	1	2	1
25	1	2	1
26	1	2	1
27	1	2	1
28	1	2	1
29	1	2	1
30	1	1	1
31	1	2	1
32	1	2	1
33	1	2	1
34	1	2	1
35	1	2	1
36	1	1	2
37	1	2	1
38	1	1	1
39	1	1	1
40	1	2	1
41	1	2	1
42	1	2	1
43	1	2	1

	Single	complete	average
44	1	1	1
45	1	2	1
46	1	2	1
47	1	2	1
48	1	2	1
49	1	2	1
50	1	2	1
51	1	1	1
52	1	2	1
53	1	1	2
54	1	1	2
55	1	1	2
56	1	1	2
57	1	1	2
58	1	1	2
59	1	2	2
60	1	1	2
61	1	1	2
62	1	1	2
63	1	1	2
64	1	1	2
65	1	1	2
66	1	1	2
67	1	1	2
68	1	2	2
69	1	1	2
70	1	1	2
71	1	2	2
72	1	1	2
73	1	1	2
74	1	1	2
75	1	1	2
76	1	1	2
77	1	1	2
78	1	2	2
79	1	1	2
80	1	2	2
81	1	1	2
82	2	1	2
83	2	2	2
84	2	2	2
85	2	1	2
86	2	2	2
87	2	2	2
88	2	1	2
89	2	2	2
90	2	1	2
91	2	2	2

	Single	complete	average
92	2	1	2
93	2	2	2
94	2	2	2
95	2	2	2
96	2	2	2
97	2	2	2
98	2	2	2
99	2	2	2
100	2	2	2
101	2	2	2
102	2	2	2
103	2	2	2
104	2	2	2
105	2	2	2
106	2	1	2
107	2	2	2

b. Analisis *cluster* yang dibentuk sama dengan tiga (k = 3)

	Single	complete	average
1	1	1	1
2	1	2	2
3	1	2	2
4	1	2	2
5	1	2	2
6	1	2	2
7	1	2	2
8	1	2	2
9	1	2	2
10	1	1	1
11	1	2	2
12	1	2	2
13	1	1	1
14	1	2	2
15	1	2	2
16	1	1	1
17	1	2	2
18	1	2	2
19	1	2	2
20	1	2	2
21	1	2	2
22	1	2	2
23	1	2	2
24	1	2	2
25	1	2	2

	Single	complete	average
26	1	2	2
27	1	2	2
28	1	2	2
29	1	2	2
30	1	1	1
31	1	2	2
32	1	2	2
33	1	2	2
34	1	2	2
35	1	2	2
36	2	1	3
37	1	2	2
38	1	1	1
39	1	1	1
40	1	2	2
41	1	2	2
42	1	2	2
43	1	2	2
44	1	1	1
45	1	2	2
46	1	2	2
47	1	2	2
48	1	2	2
49	1	2	2
50	1	2	2
51	1	1	1
52	1	2	2
53	2	1	3
54	2	1	3
55	2	1	3
56	2	1	3
57	2	1	3
58	2	1	3
59	2	3	3
60	2	1	3
61	2	1	3
62	2	1	3
63	2	1	3
64	2	1	3
65	2	1	3
66	2	1	3
67	2	1	3
68	2	3	3
69	2	1	3
70	2	1	3
71	2	3	3
72	2	1	3
73	2	1	3

	Single	complete	average
74	2	1	3
75	2	1	3
76	2	1	3
77	2	1	3
78	2	3	3
79	2	1	3
80	2	3	3
81	2	1	3
82	3	1	3
83	3	3	3
84	3	3	3
85	3	1	3
86	3	3	3
87	3	3	3
88	3	1	3
89	3	3	3
90	3	1	3
91	3	3	3
92	3	1	3
93	3	3	3
94	3	3	3
95	3	3	3
96	3	3	3
97	3	3	3
98	3	3	3
99	3	3	3
100	3	3	3
101	3	3	3
102	3	3	3
103	3	3	3
104	3	3	3
105	3	3	3
106	3	1	3
107	3	3	3

d. Analisis *cluster* yang dibentuk sama dengan empat (k = 4)

	Single	complete	average
1	1	1	1
2	1	2	2
3	1	2	2
4	1	2	2
5	1	2	2
6	1	2	2
7	1	2	2

	Single	complete	average
8	1	2	2
9	1	2	2
10	1	1	1
11	1	2	2
12	1	2	2
13	1	1	1
14	1	2	2
15	1	2	2
16	1	1	1
17	1	2	2
18	1	2	2
19	1	2	2
20	1	2	2
21	1	2	2
22	1	2	2
23	1	2	2
24	1	2	2
25	1	2	2
26	1	2	2
27	1	2	2
28	1	2	2
29	1	2	2
30	1	1	1
31	1	2	2
32	1	2	2
33	1	2	2
34	1	2	2
35	1	2	2
36	2	1	3
37	1	2	2
38	1	1	1
39	1	1	1
40	1	2	2
41	1	2	2
42	1	2	2
43	1	2	2
44	1	1	1
45	1	2	2
46	1	2	2
47	1	2	2
48	1	2	2
49	1	2	2
50	1	2	2
51	1	1	1
52	1	2	2
53	2	3	3
54	2	3	3
55	2	3	3

	Single	complete	average
56	2	3	3
57	2	3	3
58	2	3	3
59	2	4	4
60	2	3	3
61	2	3	3
62	2	3	3
63	2	3	3
64	2	3	3
65	2	3	3
66	2	1	3
67	2	1	3
68	2	4	4
69	2	3	3
70	2	1	3
71	2	4	4
72	2	3	3
73	2	1	3
74	2	3	3
75	2	3	3
76	2	3	3
77	2	3	3
78	2	4	4
79	2	3	3
80	2	4	4
81	2	3	3
82	3	3	3
83	3	4	4
84	3	4	4
85	3	3	3
86	3	4	4
87	3	4	4
88	3	3	3
89	3	4	4
90	3	3	3
91	3	4	4
92	3	3	3
93	3	4	4
94	3	4	4
95	3	4	4
96	3	4	4
97	3	4	4
98	3	4	4
99	3	4	4
100	3	4	4
101	3	4	4
102	4	4	4

	Single	complete	average
103	3	4	4
104	3	4	4
105	3	4	4
106	3	3	3
107	3	4	4

e. Analisis cluster yang dibentuk sama dengan lima (k = 5)

	Single	complete	average
1	1	1	1
2	1	2	2
3	1	2	2
4	1	2	3
5	1	2	2
6	1	2	2
7	1	3	3
8	1	2	2
9	1	2	2
10	1	1	1
11	1	3	3
12	1	2	3
13	1	1	1
14	1	2	2
15	1	2	2
16	1	1	1
17	1	2	3
18	1	2	3
19	1	2	3
20	1	3	3
21	2	3	3
22	1	2	3
23	1	2	2
24	1	2	2
25	2	3	3
26	1	2	3
27	1	3	3
28	1	3	3
29	1	3	3
30	1	1	1
31	1	2	3
32	1	2	2
33	1	3	3
34	1	2	3
35	1	2	3
36	3	1	4

	Single	complete	average
37	1	2	2
38	1	1	1
39	1	1	1
40	1	2	2
41	1	2	2
42	1	2	2
43	1	2	2
44	1	1	1
45	1	2	2
46	1	2	2
47	2	3	3
48	1	2	3
49	1	2	2
50	1	2	3
51	1	1	1
52	1	2	2
53	3	4	4
54	3	4	4
55	3	4	4
56	3	4	4
57	3	4	4
58	3	4	4
59	3	5	5
60	3	4	4
61	3	4	4
62	3	4	4
63	3	4	4
64	3	4	4
65	3	4	4
66	3	1	4
67	3	1	4
68	3	5	5
69	3	4	4
70	3	1	4
71	3	5	5
72	3	4	4
73	3	1	4
74	3	4	4
75	3	4	4
76	3	4	4
77	3	4	4
78	3	5	5
79	3	4	4
80	3	5	5
81	3	4	4
82	4	4	4
83	4	5	5

	Single	complete	average
84	4	5	5
85	4	4	4
86	4	5	5
87	4	5	5
88	4	4	4
89	4	5	5
90	4	4	4
91	4	5	5
92	4	4	4
93	4	5	5
94	4	5	5
95	4	5	5
96	4	5	5
97	4	5	5
98	4	5	5
99	4	5	5
100	4	5	5
101	4	5	5
102	5	5	5
103	4	5	5
104	4	5	5
105	4	5	5
106	4	4	4
107	4	5	5

Lampiran 9. *Syntax* Rasio *Sw dan Sb* Metode *AGNES*

```
StdNumerik1<-data.frame(StdNumerik$StdMhsIPK,StdNumerik$StdMhsSKS)
d<-dist(StdNumerik1, method = "euclidean")

# Analisis Cluster Hirarki
fit.sin = hclust(d, method = "single")
fit.com = hclust(d, method = "complete")
fit.ave = hclust(d, method = "average")

#  Memotong dendogram untuk k cluster
single = cutree(fit.sin, k=2)
complete = cutree(fit.com, k=5)
average = cutree(fit.ave, k=2)
 hasil.cluster.numerik<-data.frame(single,complete,average)
dataNumerik2<-cbind(StdNumerik$StdMhsIPK,StdNumerik$StdMhsSKS)
```

METODE *SINGLE LINKAGE*

```
# Analisis Cluster metode Single Linkage
cluster.single<-hasil.cluster.numerik$single
data.single<-data.frame(single,datarata)
data.single.sort<-data.single[order(data.single$single),]

# menghitung Sw
mean.c1.single<-mean(data.c1.single)
mean.c2.single<-mean(data.c2.single)

sw1.single<-sqrt((sum((data.c1.single-mean.c1.single)^2))/(81))
sw2.single<-sqrt((sum((data.c2.single-mean.c2.single)^2))/(26))

jumlah.sw.single<-sum(sw1.single,sw2.single)
sw.single<-jumlah.sw.single/(2)

# Menghitung nilai Sb
sb1.single<-((mean.c1.single-mean(datarata))^2)
sb2.single<-((mean.c2.single-mean(datarata))^2)

jumlah.sb.single<-sum(sb1.single,sb2.single)
sb.single<-sqrt(jumlah.sb.single/(2-1))

# Menghitung Ratio perbandingan Sw dan Sb
ratio.single<-sw.single/sb.single
hasil.single<-c(sw.single,sb.single,ratio.single)
```

METODE *COMPLETE LINKAGE*

```
# Analisis Cluster metode Complete Linkage
    cluster.complete<-hasil.cluster.numerik$complete
    data.complete<-data.frame(complete,datarata)
    data.complete.sort<-data.complete[order(data.complete$complete),]

# menghitung Sw
    mean.c1.complete<-mean(data.c1.complete)
    mean.c2.complete<-mean(data.c2.complete)
    mean.c3.complete<-mean(data.c3.complete)
    mean.c4.complete<-mean(data.c4.complete)
    mean.c5.complete<-mean(data.c5.complete)

    sw1.complete<-sqrt((sum((data.c1.complete-mean.c1.complete)^2))/(14))
    sw2.complete<-sqrt((sum((data.c2.complete-mean.c2.complete)^2))/(32))
    sw3.complete<-sqrt((sum((data.c3.complete-mean.c3.complete)^2))/(10))
    sw4.complete<-sqrt((sum((data.c4.complete-mean.c4.complete)^2))/(26))
    sw5.complete<-sqrt((sum((data.c5.complete-mean.c5.complete)^2))/(25))

    jumlah.sw.complete<-
        sum(sw1.complete,sw2.complete,sw3.complete,sw4.complete,sw5.compl
        ete)
    sw.complete<-jumlah.sw.complete/(5)


# Menghitung nilai Sb
    sb1.complete<-((mean.c1.complete-mean(datarata))^2)
    sb2.complete<-((mean.c2.complete-mean(datarata))^2)
    sb3.complete<-((mean.c3.complete-mean(datarata))^2)
    sb4.complete<-((mean.c4.complete-mean(datarata))^2)
    sb5.complete<-((mean.c5.complete-mean(datarata))^2)

    jumlah.sb.complete<-
        sum(sb1.complete,sb2.complete,sb3.complete,sb4.complete,sb5.complete,
    sb.complete<-sqrt(jumlah.sb.complete/(5-1))

# Menghitung Ratio perbandingan Sw dan Sb
    ratio.complete<-sw.complete/sb.complete
    hasil.complete<-c(sw.complete,sb.complete,ratio.complete)
```

METODE *AVERAGE LINKAGE*

```
# Analisis Cluster metode Complete Linkage
    cluster.average<-hasil.cluster.numerik$average
    data.average<-data.frame(average,datarata)
    data.average.sort<-data.average[order(data.average$average),]

# menghitung Sw
    mean.c1.average<-mean(data.c1.average)
    mean.c2.average<-mean(data.c2.average)

    sw1.average<-sqrt((sum((data.c1.average-mean.c1.average)^2))/(51))
    sw2.average<-sqrt((sum((data.c2.average-mean.c2.average)^2))/(56))

    jumlah.sw.average<-sum(sw1.average,sw2.average)
    sw.average<-jumlah.sw.average/(2)

# Menghitung nilai Sb
    sb1.average<-((mean.c1.average-mean(datarata))^2)
    sb2.average<-((mean.c2.average-mean(datarata))^2)

    jumlah.sb.average<-sum(sb1.average,sb2.average)
    sb.average<-sqrt(jumlah.sb.average/(k-1))

#  Menghitung Ratio perbandingan Sw dan Sb
    ratio.average<-sw.average/sb.average
    hasil.average<-c(sw.average,sb.average,ratio.average)
```

Lampiran 10. *Syntax* Metode *Ensemble ROCK* untuk data Campuran

```
dataNumerik<-data.frame(DataMhs$IPK,DataMhs$SKS)
StdMhsIPK<-scale(DataMhs$IPK, center = TRUE, scale = TRUE)
StdMhsSKS<-scale(DataMhs$SKS, center = TRUE, scale = TRUE)

dn<-data.frame(StdMhsIPK,StdMhsSKS)
dk<-data.frame(DataMhs$`asal sekolah`,DataMhs$`status
      keorganisasian`,DataMhs$`pekerjaan orangtua`,DataMhs$`pendidikan
      terakhir orangtua`)

# Metode AGNES
d<-dist(dn,method = "euclidean")
fit<-hclust(d,method = "complete")
complete<-cutree(fit,k=5)

# Metode ROCK
a<-dummy.data.frame(dk)
set.seed(2017)
rc<-rockCluster(a,n=3,theta = 0,01,debug = FALSE)
rf.hasil<-fitted(rc)
theta.01<-rf.hasil$cl

# Metode Ensemble ROCK
daka<-data.frame(complete,theta.01)
ddu<-dummy.data.frame(daka)

rc.01<-rockCluster(ddu,n=3,theta = 0.01,debug = FALSE)
rc.05<-rockCluster(ddu,n=3,theta = 0.05,debug = FALSE)
rc.10<-rockCluster(ddu,n=3,theta = 0.10,debug = FALSE)
rc.25<-rockCluster(ddu,n=3,theta = 0.25,debug = FALSE)
rc.50<-rockCluster(ddu,n=3,theta = 0.50,debug = FALSE)
rc.75<-rockCluster(ddu,n=3,theta = 0.75,debug = FALSE)
rc.80<-rockCluster(ddu,n=3,theta = 0.80,debug = FALSE)
rc.95<-rockCluster(ddu,n=3,theta = 0.95,debug = FALSE)



rf.01<-fitted(rc.01)
rf.05<-fitted(rc.05)
rf.10<-fitted(rc.10)
rf.25<-fitted(rc.25)
rf.50<-fitted(rc.50)
rf.75<-fitted(rc.75)
rf.80<-fitted(rc.80)
rf.95<-fitted(rc.95)
```

```
theta.01<-rf.01$cl
theta.05<-rf.05$cl
theta.10<-rf.10$cl
theta.25<-rf.25$cl
theta.50<-rf.50$cl
theta.75<-rf.75$cl
theta.80<-rf.80$cl
theta.95<-rf.95$cl

cluster<-
    data.frame(theta.01,theta.05,theta.10,theta.25,theta.50,theta.75,theta.80,thet
    a.95)
hasil<-data.frame(cluster)
```

Lampiran 11. Output Hasil Metode *ensemble ROCK* untuk Data Campuran

	theta.01	theta.05	theta.10	theta.25
1	3	3	2	2
2	2	3	2	2
3	2	3	2	3
4	2	2	3	3
5	2	2	2	2
6	3	3	2	2
7	2	2	2	2
8	3	2	3	2
9	3	2	2	2
10	2	3	3	2
11	2	3	3	3
12	2	3	2	2
13	2	2	2	2
14	3	3	3	2
15	2	2	2	2
16	3	3	2	3
17	2	3	2	3
18	2	3	2	3
19	2	3	2	2
20	3	2	2	3
21	2	3	3	3
22	3	2	3	3
23	3	3	3	2
24	3	2	2	3
25	2	2	2	2
26	2	2	3	2
27	2	3	3	3
28	3	3	3	3
29	2	2	2	3
30	3	3	2	3
31	2	2	2	3
32	2	3	3	2
33	2	2	2	3
34	3	2	2	2
35	3	2	2	2
36	2	3	3	3
37	2	3	3	2
38	2	3	3	3
39	2	3	2	3
40	2	2	2	3
41	3	2	3	3
42	3	2	2	2
43	3	3	3	3

	theta.01	theta.05	theta.10	theta.25
44	3	2	3	3
45	2	2	2	2
46	2	3	3	3
47	3	2	2	2
48	2	3	3	2
49	3	3	2	3
50	3	3	3	3
51	2	3	3	3
52	3	2	2	3
53	2	2	3	3
54	2	3	3	2
55	2	2	2	3
56	3	2	2	3
57	2	2	2	2
58	2	3	3	2
59	3	3	2	3
60	3	3	3	2
61	3	3	2	3
62	3	2	3	3
63	2	2	3	3
64	3	2	2	3
65	3	2	2	3
66	3	2	2	2
67	3	2	2	2
68	3	3	3	2
69	2	3	3	2
70	2	2	2	2
71	2	2	2	2
72	2	2	3	2
73	3	3	3	2
74	2	3	3	3
75	2	2	2	3
76	3	3	3	3
77	2	3	3	2
78	2	3	2	2
79	3	2	3	2
80	2	3	3	2
81	3	2	3	2
82	2	2	3	3
83	2	3	3	2
84	3	3	2	3
85	3	3	3	2
86	2	3	3	3
87	2	2	3	3
88	2	2	2	3

```
     theta.01     theta.05    theta.10     theta.25
89          2            2           2            2
90          2            2           3            3
91          3            3           2            2
92          2            2           3            2
93          2            2           3            2
94          3            2           2            3
95          3            3           2            3
96          3            3           3            3
97          3            2           2            2
98          2            3           3            2
99          3            3           3            2
100         3            3           3            2
101         2            2           3            2
102         2            2           2            2
103         2            3           2            2
104         2            2           3            2
105         2            3           3            2
106         3            2           3            3
107         2            2           3            3
```

```
    theta.01     freq
1          2       61
2          3       46

    theta.05     freq
1          2       51
2          3       56

    theta.10     freq
1          1       54
2          2       53

    theta.25     freq
1          1       54
2          2       53
```

	theta.50	theta.75	theta.80	theta.95
1	1	1	1	1
2	1	1	1	1
3	2	2	2	2
4	2	2	2	2
5	1	1	1	1
6	1	1	1	1
7	2	2	2	2
8	2	2	2	2
9	2	2	2	2
10	2	2	2	2
11	1	1	1	1
12	2	2	2	2
13	2	2	2	2
14	2	2	2	2
15	2	2	2	2
16	2	2	2	2
17	2	2	2	2
18	1	1	1	1
19	1	1	1	1
20	1	1	1	1
21	1	1	1	1
22	2	2	2	2
23	1	1	1	1
24	1	1	1	1
25	2	2	2	2
26	1	1	1	1
27	1	1	1	1
28	2	2	2	2
29	2	2	2	2
30	2	2	2	2
31	1	1	1	1
32	2	2	2	2
33	2	2	2	2
34	2	2	2	2
35	1	1	1	1
36	1	1	1	1
37	2	2	2	2
38	2	2	2	2
39	2	2	2	2
40	1	1	1	1
41	1	1	1	1
42	1	1	1	1
43	2	2	2	2
44	3	2	2	2
45	1	1	1	1

	theta.50	theta.75	theta.80	theta.95
46	1	1	1	1
47	2	2	2	2
48	1	1	1	1
49	1	1	1	1
50	1	1	1	1
51	1	1	1	1
52	1	1	2	2
53	2	2	1	1
54	1	1	1	1
55	1	1	1	2
56	2	2	2	2
57	2	2	1	1
58	1	1	1	1
59	1	1	1	1
60	1	1	2	2
61	2	2	2	2
62	2	2	1	1
63	1	1	1	1
64	1	1	1	1
65	1	1	2	2
66	2	2	2	2
67	2	2	2	2
68	2	2	1	1
69	1	1	1	1
70	1	1	1	1
71	1	1	1	1
72	1	1	2	2
73	2	2	1	1
74	1	1	1	1
75	1	1	2	2
76	2	2	2	2
77	2	2	1	1
78	1	1	2	2
79	2	2	1	1
80	1	1	2	2
81	2	2	1	1
82	1	1	1	1
83	1	1	1	1
84	1	1	2	2
85	2	2	2	2
86	2	2	2	2
87	2	2	2	2
89	2	2	2	2
90	2	2	2	2

```
     theta.50     theta.75     theta.80     theta.95

91          1            1            1            1
92          1            1            1            1
93          1            1            1            1
94          1            1            1            1
95          1            1            1            1
96          2            2            2            2
97          1            1            1            1
98          2            2            2            2
99          1            1            1            1
100         2            2            2            2
101         1            1            1            1
102         1            1            1            1
103         2            2            2            2
104         2            2            2            2
105         1            1            1            1
106         2            2            2            2
107         2            2            2            2


 ------------------------------------------------------------------------------------------

    theta.50     freq
1          2       57
2          3       50

    theta.75     freq
1          2       51
2          3       56

    theta.80     freq
1          1       54
2          2       53

    theta.95     freq
1          1       54
2          2       53
```

RIWAYAT HIDUP

Nur Ariska, lahir di Pinrang pada tanggal 4 September 1995, Anak ke Dua dari Lima bersaudara. buah hati dari pasangan Anas dan Hasni.

Mulai memasuki jenjang pendidikan Sekolah Dasar pada tahun 2001 di SDN 178 Lanrisang, Pinrang dan tamat pada tahun 2007 di SDN 15 Kotu, Enrekang. Pada tahun 2007 melanjutkan Pendidikan di SMP Negeri 3 Anggeraja, Enrekang dan tamat pada tahun 2010. Kemudian pada tahun yang sama melanjutkan pendidikan di SMA Negeri 1 Anggeraja, Enrekang dan tamat tahun 2013. Pada tahun 2013, penulis melanjutkan pendidikan di di Program Studi Statistika (S1) Fakultas Matematika dan Ilmu Pengetahuan Alam Universitas Negeri Makassar. Selama menjalani akademik, penulis terlibat orgnisasi dalam kampus yaitu, HIMASTAT FMIPA UNM periode 2013-2015 dan periode 2015-2016. Penulis dapat dihubungi melalui email ariskanur1@gmail.com.